DOCUMENT RESUME

ED 406 392 TM 026 255

TITLE	North Carolina End-of-Course Tests: Algebra I; Biology; Economic, Legal, and Political Systems; English I; U.S. History. Technical Report No. 1.
INSTITUTION	North Carolina State Dept. of Public Instruction, Raleigh. Div. of Accountability/Testing.
PUB DATE	Sep 96
NOTE	139p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Algebra; *Biology; Citizenship Education; *Civics; *English; High Schools; Law Related Education; Psychometrics; Scaling; Scores; State Programs; *Test Construction; Testing Programs; Test Items; Test Reliability; Test Results; Test Use; Test Validity; United States History
IDENTIFIERS	Exit Examinations; *North Carolina End of Course Testing Program

ABSTRACT

The North Carolina End-of-Course Testing Program is based on the assessment of higher level skills in the context of specific subject-area content. These tests inform students, parents, the community, and educators about the achievement of North Carolina students in given areas. This report describes the development and psychometric properties of the end-of-course tests in Algebra I; Biology; Economic, Legal, and Political Systems; English I; and United States History. The value of these tests lies primarily in the fact that the scores provide a common yardstick that is not influenced by local differences. The end-of-course test score counts as part of the student's grade in the course, with how much it counts being a local decision. The tests described in this publication are administered during the last 10 days of a course. For each of these tests, the report describes: (1) item development; (2) test development; (3) scores and scales; (4) descriptive statistics and reliability; and (5) validity. Six appendixes provide sample test items and discuss aspects of test construction and scoring in detail. (Contains 19 tables, 11 figures, 19 appendix graphs, 4 appendix tables, and 34 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

Technical Report #1

ED 406 392

NORTH CAROLINA

End-of-Course Tests

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

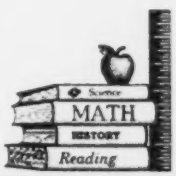
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

E. BRUMBACK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Algebra I
Biology
Economic, Legal, & Political Systems
English I
U.S. History



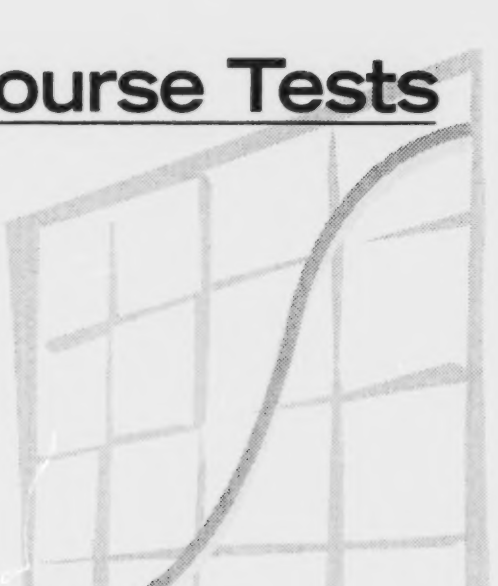
NCTests

TM026255

Public Schools of North Carolina
State Board of Education
Department of Public Instruction
Office of Instructional and Accountability Services
Division of Accountability/Testing

BEST COPY AVAILABLE

End-of-Course Tests



Algebra I
Biology
Economic, Legal, & Political Systems
English I
U.S. History

Prepared by

Eleanor E. Sanford, Ph.D.
September 1996

Technical Support	Laura Kramer, Jeff Morgan, Nada Ballator, and Lori D. McLeod
	L.L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill: David Thissen, Val S.L. Williams, Mary Pommerich, Kathy Billeaud, Lori McLeod, and Lauren Nelson.
Test Production	Shirley Stoll, Susan Godwin, Sandra B. Mills, Trena Brantley, Adrienne Silvay, Patricia Atkinson
Others	Chris Averett, Suzanne Triplett, William J. Brown, Bob Evans, Daisy Vickers, Mildred Bazemore, Gary Williamson, and Doris Tyler

For additional information contact

Eleanor E. Sanford, Ph.D.
NCDPI/Testing
301 North Wilmington Street
Raleigh, North Carolina 27601-2825
919/715-1214
e-mail: esanford@dpi.state.nc.us

Table of Contents

List of Tables ii
List of Figures iii

Introduction 1
Background 1
Related Testing Materials 3

Description of Tests 5
Algebra I 6
Biology 7
Economic, Legal, and Political Systems (ELP) 7
English I 7
US History 8

Item Development 11
Content Development 11
Test Specifications 14
Item Writing and Review 23
Field Testing 25
Item Analysis and Selection 27

Test Development 35

Scores and Scales 39
Scales 39
Scores 39

Descriptive Statistics and Reliability 41
Descriptive Statistics 41
Reliability 44

Validity 47
Content Validity 47
Criterion-Related Validity 47
Construct Validity 50

Resources 55

Appendices 57

List of Tables

1. North Carolina End-of-Course Testing Program 2

2. Administrative information for the North Carolina End-of-Course Tests 5

3. End-of-Course Test Development Calender 12

4. Item pool specifications for the North Carolina End-of-Course Test of Algebra I—Percent of items on the test assessing each goal of the curriculum 17

5. Item pool specifications for the North Carolina End-of-Course Test of ELP—Percent of items on the test assessing each goal of the curriculum 19

6. English II Conventions Scoring Criteria as applied to the editing item pool of English I 20

7. Item pool specifications for the North Carolina End-of-Course Test of US History—Percent of items on the test assessing each goal of the curriculum 22

8. Number of items developed by subject area 24

9. Number of items and passages field tested in each subject area 25

10. Characteristics of the field test samples 26

11. Average item pool parameter estimates for the North Carolina End-of-Course Tests 32

12. Average p-value for the total test and each subtest of the North Carolina End-of-Course Tests 35

13. Average test review ratings for the North Carolina End-of-Course Tests 37

14. Descriptive statistics for the first administration of the North Carolina End-of-Course Tests—Forms A, B, and C 41

15. Item-level values of coefficient α for the 1993 administration of the North Carolina End-of-Course Tests (Forms A, B, and C) and pretests 44

16. Standard error of measurement for ranges of scores on the North Carolina End-of-Course Tests 45

17. Percent of students assigned to each achievement level by teachers during field testing 48

18. Range of scores associated with each achievement level for score reporting 50

19. Results from Survey Question on Laboratory Experience and the North Carolina Field Test of Biology 53

List of Figures

1. Test development process for the North Carolina End-of-Course Test of Algebra I 10

2. Thinking skills framework used with the North Carolina End-of-Course Tests (*adapted from Robert Marzano et al, *Dimensions of Thinking,* 1988) 15

3. Item characteristic curve of a typical 4-option multiple-choice item (a = 1.00, b = 0.00, and c = 0.20) 28

4. Item characteristic curve of an item that exhibited a low slope (Algebra I item #2098, a = 0.357, b = 0168, and c = 0.229). This item was flagged as exhibiting "Weak Prediction" 29

5. Item characteristic curve of an item that was difficult, but was retained for test development (Biology item #4106, a = 1.298, b = 2.499, and c = 0.089) 29

6. Frequency distribution of scores on the North Carolina End-of-Course Test of Algebra I—Forms A, B, and C (N = 69,162) 42

7. Frequency distribution of scores on the North Carolina End-of-Course Test of Biology—Forms A, B, and C (N = 62,470) 42

8. Frequency distribution of scores on the North Carolina End-of-Course Test of English I—Forms A, B, and C (N = 67,739) 43

9. Frequency distribution of scores on the North Carolina End-of-Course Test of US History—Forms A, B, and C (N = 53,157) 43

10. The relationship between teacher judgments of student achievement and scores on the North Carolina End-of-Course Tests during field testing 49

11. Comparison of letter grades teachers expected students to receive and scores subsequently earned on the North Carolina End-of-Course Tests 52

During this decade and for many decades to come, North Carolina students will need to move far beyond the mastery of basic skills to the mastery of higher level skills. The term "higher level skills" refers to the thinking and problem solving strategies that enable people to access, sort, and digest enormous amounts of information. It refers to the skills required to solve complex problems and to make informed choices and decisions. It also refers to advanced communication skills that enable individuals to express and share what they know and to work well with others (North Carolina End-of-Grade Testing Program: Background Information, 1993, p. 1).

The End-of-Course Testing Program is based on the assessment of these higher level skills within the context of specific subject-area content. When properly administered and interpreted, these test results provide an independent, uniform source of reliable and valid information which enables

- *students* to know the extent to which they have mastered expected knowledge and skills and how they compare to others;
- *parents* to know if their children are acquiring the knowledge and skills needed to succeed in a highly competitive job market;
- *teachers* to know if their students have mastered grade-level knowledge and skills in the curriculum and, if not, what weaknesses need to be addressed;
- *community leaders and lawmakers* to know if students in North Carolina schools are improving their performance over time and how the students compare with students from other states or the nation; and
- *citizens* to objectively assess their return on investment in the public schools

(*North Carolina Testing Code of Ethics*, revised 1996).

This technical report describes the development and psychometric properties of the North Carolina End-of-Course Tests of Algebra I; Biology; Economic, Legal, and Political Systems; English I; and United States History.

Background

The North Carolina End-of-Course Testing Program was initiated in response to legislation passed by the North Carolina General Assembly—the North Carolina Elementary and Secondary Reform Act of 1984. This act called for the implementation of the Basic Education Program through the establishment of a core curriculum for all students for each content area and the development of tests to assess the implementation of each curriculum across the state. The following sections of the *Public School Laws* (1994) describe the legislation in relation to the development of tests.

> Public School Law 115C-174.10 defines the following "purposes of testing programs in North Carolina: (1) to assure that all high school graduates possess the . . . skills and knowledge thought necessary to function as a member of society; (2) to provide a means of identifying strengths and weaknesses in the education process; and (3) to establish additional means for making the education system accountable to the public for results."
>
> Public School Law 115C-174.11 (c) calls for the "adoption of a system of end-of-course testing designed to measure progress toward selected competencies, especially core academic competencies, described in the *Standard Course of Study* for appropriate grade levels. With regard to students who are identified as not demonstrating satisfactory academic progress, end-of-course . . . test results shall be used in developing strategies and plans for assisting those students in achieving satisfactory academic progress."

Based on these statutes, the North Carolina End-of-Course Testing Program was developed for two purposes:

- to provide accurate measurement of individual student skills and knowledge specified in the North Carolina *Standard Course of Study*, and
- to provide accurate measurement of the knowledge and skills attained by groups of students for school, school system, and state accountability.

Scores on the end-of-course tests are only one of many indicators of the achievement of students. The value of these tests lies primarily in the fact that the scores provide a common standard that is not influenced by local differences in achievement and expectations. The tests provide yardsticks which can be used to compare the achievement of students, schools, school systems, and the state. The assessment yardstick can be used to measure gains (or losses) in performance across time to see if educational improvement efforts at the state and local level are working.

The North Carolina End-of-Course Testing Program includes multiple choice assessments of Algebra I; Biology; Economic, Legal, and Political Systems (ELP); English I; and US History. Writing is assessed by the English II test administered to students in grade 10 during March. Each student completes two prompts assessing two of the four modes of writing—expository, descriptive, narrative, and argumentative. All of the tests are developed by the North Carolina Department of Public Instruction and are aligned with the revised North Carolina *Standard Course of Study*.

Table 1. North Carolina End-of-Course Testing Program.

	English Language Arts	Mathematics	Social Studies	Science	Competency
9	**English I**	**Algebra I**	**ELP**	Physical Science	
10	**English II**	Geometry		**Biology**	**Reading & Math**
11		Algebra II	US History	Chemistry	**Reading & Math**
12				Physics	**Reading & Math**

Note: Shaded blocks are state-mandated tests; unshaded blocks are local option tests.

In response to Senate Bill 16 passed in 1995, the State Board of Education examined the structure and functions of the state public school system in order to improve student performance, increase local flexibility and control, and promote economy and efficiency. In May 1995, the State Board of Education issued *The New ABCs of Public Education: Accountability, curriculum Basics, and local Control and flexibility*. One of the key recommendations was to adopt a new accountability plan focused on the performance of individual public schools (rather than school systems) on the basic courses required for graduation from high school. Rather than comparing different students over time, this plan—the School-based Management and Accountability Program (legislation passes during the 1996 session of the North Carolina General Assembly)—will hold schools accountable for the educational growth of groups of students over time.

For school, school system, and state accountability, the multiple-choice Algebra I, Biology, ELP, English I, and US History tests will be used to monitor growth. One proposal for monitoring growth at the high school level is to use a pretest/posttest design; the difference between scores from the pretest and the end-of-course test at the end of the course will be used to estimate growth in achievement in the subject area. "The state will set standards for growth for the given amount of schooling, expecting at least a year's worth of growth for a year's worth of school" (North Carolina State Board of Education, 1996).

For student accountability, the State Board of Education has stated that a student's end-of-course test score shall count as part of his or her overall grade in the course (it is a local decision as to how much the end-of-course test score counts in the overall grade). In addition, the grade 8 end-of-grade tests are used as a way for students to demonstrate that they have the knowledge and skills necessary to meet the reading and mathematics competency requirement for high school graduation. If a student does not score at or above Level III on the grade 8 tests then he or she must take the North Carolina Competency Tests of Reading Comprehension and/or Mathematics during high school.

Related Testing Materials

The following materials are being developed for use in conjunction with the North Carolina End-of-Course Testing Program.

Testlets The End-of-Course Testlets for Algebra I, ELP, English I, and US History (Edition A) will consist of mini-tests containing both multiple-choice and open-ended items designed to assist teachers in formative assessment (the ongoing assessment of students strengths and weaknesses). For English I, each testlet will be based on the difficulty level of the items (easy, medium, or hard items) and will contain two or three literary passages and two or three editing passages. For the other three sets of testlets, each will be related to a specific goal and objective of the North Carolina *Standard Course of Study* for the specific subject area. The testlets for Edition A will be developed from items field tested in May 1996 and then distributed in 1997.

Local Option Tests As part of the restructuring of the North Carolina public schools in response to Senate Bill 16 in 1995, the number of state-mandated tests was cut in half to focus on the basics. The State Board of Education directed the NCDPI to develop procedures to support and facilitate the continued use of non-state-mandated tests and related services by local schools and systems. School systems have been given the option to purchase previously- or newly-developed NCDPI tests for local accountability. These assessments were designed to assess Geometry, Algebra II, Physical Science, Chemistry, and Physics. These subject areas are part of the North Carolina *Standard Course of Study* and must be taught.

The North Carolina End-of-Course Tests were developed by the North Carolina Department of Public Instruction in conjunction with the L.L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill and the North Carolina Technical Advisory Group (see Appendix C for members of the group). The tests were developed for use as achievement tests to measure the acquistion of specific subject-area content knowledge and skills. The purpose of these tests is twofold: (1) to improve student performance on the knowledge and skills specified in the North Carolina *Standard Course of Study*; and (2) to hold schools, school systems, and the state accountable for the education of students on the knowledge and skills specified in the North Carolina *Standard Course of Study*. Both norm-referenced (where the frame of reference is a specified population of students) and criterion-referenced (where the frame of reference is a specified content domain) interpretations of the test scores support the purpose of the North Carolina End-of-Course Tests.

The end-of-course tests are aligned with the revised North Carolina *Standard Course of Study* and emphasize higher level thinking skills—students are expected to have knowledge of important ideas and concepts; understand and interpret events; apply knowledge, skills, and concepts; and make connections. While knowledge of facts and concepts is important, the questions on the tests are typically at a broader level and concern major ideas that students are expected to know to be considered literate. In addition to being asked to solve problems, students are asked "how" to solve a problem or "what strategy should be used" to solve a problem. Better students are able to take responsibility for their own learning. They develop an awareness of their own thinking, including attitudes, habits, and dispositions.

The North Carolina End-of-Course Tests of Algebra I, Biology, ELP, English I, and US History are administered during the last ten days of the course. While the tests are designed to assess subject area skills and knowledge associated with the specific course, other content areas are integrated into the assessments—the textual analysis part of the English I test includes social studies passages and the reading and interpretation of graphs

Table 2. Administrative information for the North Carolina End-of-Course Tests.

Subject Area		Amount of Testing Time	Number of Items on Each Form	Number of Items for Curricular Assessment
Algebra I	Pretest	55	25/15	
	EOC Test	110	51/30	243
Biology	Pretest	55	44	
	EOC Test	110	95	285
ELP	Pretest	55	50	
	EOC Test	110	100	300
English I	Pretest	55	15/20	
	EOC Test	110	30/42	216
US History	Pretest	55	50	
	EOC Test	110	100	300

and charts; the Algebra I test incorporates science and social studies data and experiments as sources of data; the Biology test incorporates data analysis and statistics (e.g., probability); and the US History test incorporates literature through the analysis of the impact of authors and their works on the political and social events of the time. During the 1996-97 school year, pretests will be administered on a voluntary basis to all students within the first two to three weeks of the school year (or course if block scheduling is used) to develop a database of information concerning the growth of student achievement in each of the subject areas.

All of the tests are scanned and scored within the local school system. Individual reports and school and school system summary reports are printed at the local level for accountability (the *Report Card*, the *Administrative Supplement*, and School-Building Improvement Reports).

For security and curricular purposes, multiple test forms are administered in each classroom. The measurement of student achievement is attained by administering different sets of items (equivalent in difficulty) to all students. The assessment of curriculum is met by the sum of the items administered across the three forms. All forms are administered in each classroom, one form per student. The measurement afforded by the three forms of items is critical to assessing curriculum mastery at the classroom, school, school system, and state levels.

Algebra I

The North Carolina End-of-Course Test and Pretest of Algebra I assesses three broad areas of the curriculum:

- basic operations, equations, and inequalities;
- functions and graphing; and
- polynomials and non-linear equations.

Both the EOC test and the pretest consist of two parts defined by the use of a calculator; students are expected to have access to a graphing calculator during Part 2 of the test. At the student level, the two parts of the test are combined to produce one score. Both the test and the pretest are aligned with the North Carolina *Standard Course of Study* for Algebra I and scores are reported in terms of scale scores, percentiles, and achievement levels.

Part 1 of the test focuses on the student's knowledge of specific procedures. For example, the students are expected to solve systems of equations by using the addition-and-subtraction method. The focus of the items written for this objective are on the application of the particular method to solve the problems, not just the solution.

Part 2 of the test focuses on problem-solving and assesses a student's ability to apply mathematical principles, solve problems, and explain mathematical processes. Problems are typically posed as real situations that students may have encountered, but are not always simple in nature (e.g., quadratic or exponential funtions in science). Students are allowed to use calculators on this part of the test. These items tend to require more reading than found on typical multiple-choice tests of mathematics.

Examples of typical items can be found in Appendix A.

Biology

The North Carolina End-of-Course Test and Pretest of Biology are aligned with the North Carolina *Standard Course of Study* for Biology and scores are reported in terms of scale scores, percentiles, and achievement levels.

On the test, students are expected to have knowledge of important principles and concepts, understand and interpret laboratory activities, and relate scientific information to everyday situations. Each item is related to one of the content objectives and, in addition, each item is categorized according to the skill goal that is emphasized. The content objectives (goals 6, 7, and 8) of the Biology *Standard Course of Study* describe the knowledge that is to be taught in Biology classes in North Carolina and determine the content of the items on the test. The skills goals—nature of science (goal 1), process skills (goals 2), manipulative skills (goal 3), societial and environmental issues (goal 5), and conceptual knowledge—are the basis for scientific literacy.

Examples of typical items can be found in Appendix A.

Economic, Legal, and Political Systems (ELP)

The North Carolina End-of-Course Test and Pretest of Economic, Legal, and Political Systems (ELP) are aligned with the North Carolina *Standard Course of Study* for ELP and scores are reported in terms of scale scores, percentiles, and achievement levels.

The ELP test focuses on two broad areas: (1) law and government and (2) economics. On the test, students are expected to have knowledge of important ideas and concepts, understand and interpret events in history, and connect historical people and events to current events and issues. Some questions ask the students to analyze primary and secondary source documents, charts and graphs, and political cartoons. Students are expected to gain a practical understanding of those systems that affect their lives as consumers and young adult citizens.

Examples of typical items can be found in Appendix A.

English I

The North Carolina End-of-Course Test and Pretest of English I consists of two parts—editing and textual analysis. At the student level, the two parts of the test are combined to produce one score. Both the test and the pretest are aligned with the North Carolina *Standard Course of Study* for English I and scores are reported in terms of scale scores, percentiles, and achievement levels.

Part 1 of the test focuses on editing and revising for English conventions. This part of the test is presented as a "peer editing" task (students are told that the passages represent drafts of student essays); students are asked to edit short essays for sentence formation, usage, mechanics, and spelling. Words or phrases that may or may not contain an error are underlined and students are asked to choose from three possible corrections or "Make no change."

Part 2 of the test focuses on textual analysis. On this part of the test, students are asked to read four to five passages representing various genres, including literary, informational, and practical passages. Some of the passages may be visual "texts" such as political cartoons, maps, works of art, or advertisements. Then the students are asked to answer questions in which they must analyze the text using literary terms and elements. Some items also assess a student's ability to read and comprehend written material that is appropriate in terms of difficulty and content. The test assesses a student's ability to use strategies which enhance reading comprehension including acquiring, interpreting, and applying information; and reading for critical analysis and evaluation.

The reading passages on the test are chosen to reflect the variety of reading done by students in and out of the classroom. The passages tend to be longer and more complete (compared to those typically found on standardized achievement tests) and have a high interest level for students. The variety of passages on each form allows for the assessment of reading for various purposes: for literary experience, to gain information, and to perform a task.

Examples of the passages and associated items can be found in Appendix A.

US History

The North Carolina End-of-Course Test and Pretest of US History are aligned with the North Carolina *Standard Course of Study* for US History and scores are reported in terms of scale scores, percentiles, and achievement levels.

On the tests students are expected to have knowledge of important ideas and concepts, understand and interprete events in history, and connect historical people and events across time. Some of the items ask the students to anlayze primary and secondary source documents. In a US History course students learn about the economic, social, and political events that have impacted our nation during the past three centuries and understand how each of these trends have shaped our nation today.

Examples of typical items can be found in Appendix A.

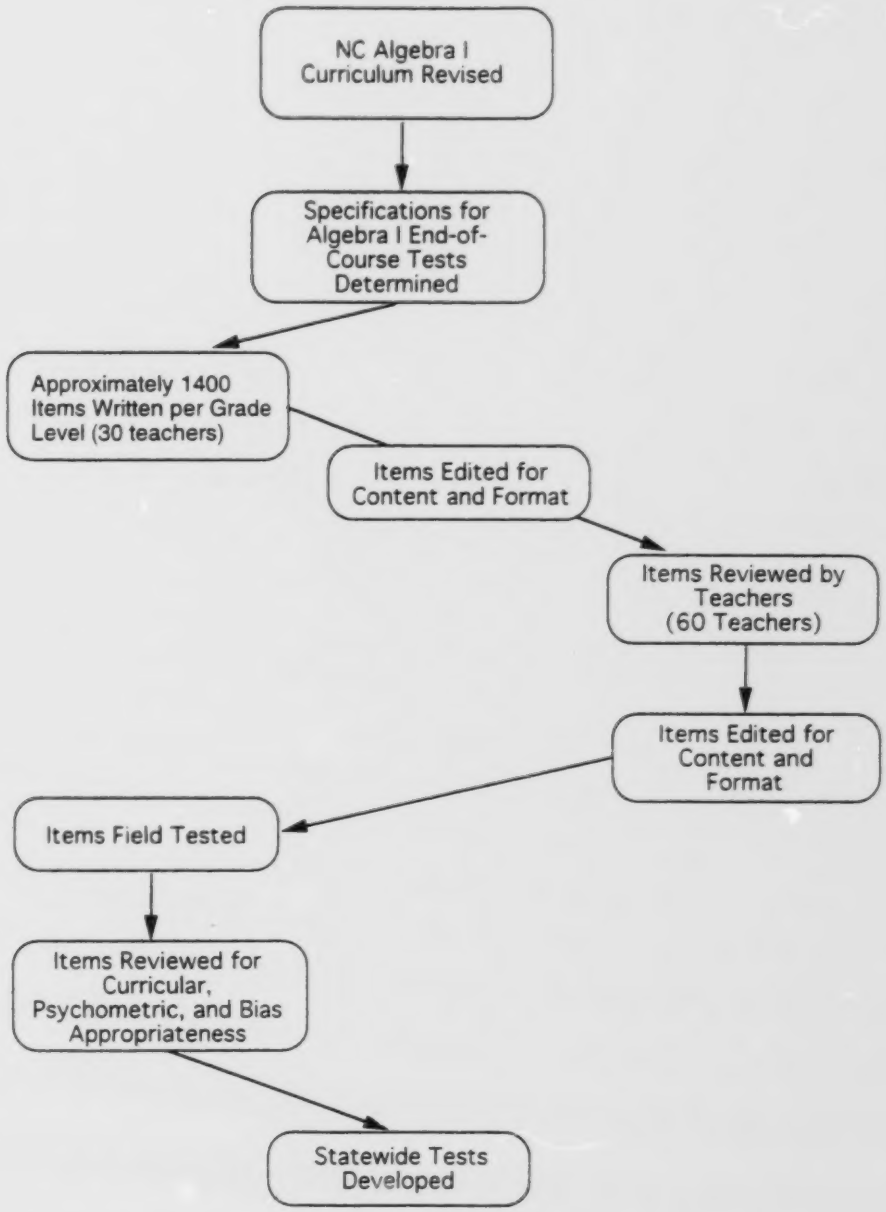

NC Algebra I Curriculum Revised
Specifications for Algebra I End-of-Course Tests Determined
Approximately 1400 Items Written per Grade Level (30 teachers)
Items Edited for Content and Format
Items Reviewed by Teachers (60 Teachers)
Items Edited for Content and Format
Items Field Tested
Items Reviewed for Curricular, Psychometric, and Bias Appropriateness
Statewide Tests Developed

Figure 1. Test development process for the North Carolina End-of-Course Test of Algebra I.

Item development for the North Carolina End-of-Course Tests goes through several unique stages (see flow chart to left for Algebra I)—content and test specification, item writing and review, field testing and analyses, and final evaluation. Each of these stages will be described in detail in the following sections. Each of the tests—Algebra I, Biology, ELP, English I, and US History—underwent the same development process, but on a different schedule. The discussion of each of these stages that follows will treat the tests as if they were developed at the same time. The chart on the next page identifies the actual time-frame associated with each stage for each test.

Content Development

If a test is to be used to measure the degree to which a course of study has been mastered, the first step is to define the curriculum. The curricula assessed by the North Carolina End-of-Course Tests were developed by the NCDPI/Division of Program Services, involving curriculum specialists, teachers, administrators, university professors, and others. These curricula reflect national content standards for student performance and the educational requirements in other industrialized nations. Content validity—the degree to which test items reflect the basic instructional program—is a quality commonly referenced in evaluating achievement tests. Content validity is built into a test from the beginning, and the procedures related to the content validation of the North Carolina End-of-Course Tests are described below.

Mathematics The North Carolina Mathematics curriculum is closely aligned with the National Council of Teachers of Mathematics *Curriculum and Evaluation Standards for School Mathematics* (1989) guidelines for teaching mathematics in kindergarten through grade 12 . These standards describe "mathematics instruction that must adopt broader goals for*all* students. It must provide experiences that encourage and enable students to value mathematics, gain confidence in their own mathematical ability, become mathematical problem solvers, communicate mathematically, and reason mathematically" (page 123). The *Teacher Handbook for Mathematics* (NCDPI, 1992, page 3) states that

> the mathematics program proposed here is by necessity broader and more inclusive than in the past. It must develop more than vocabulary, facts, and principles; more than the ability to analyze a problem situation; more than an understanding of the logical structure of mathematics. The mathematics program must provide students with the knowledge which will enable them to distinguish fact from opinion, relevant from irrelevant data, and experimental results from proven theorems. . . . It must develop reading skills, motivation, and study habits essential for independent learning of mathematics. It must develop in students the appreciation for the connections between various branches of mathematics and how mathematics is connected to other disciplines.

English Language Arts The North Carolina reading curriculum is based on national trends in reading instruction, such as the International Reading Association, that see reading as the process of constructing meaning through the dynamic interaction among the reader's existing knowledge, the information suggested by the written language, and the context of the reading situation (Michigan Curriculum Review Committee/ Michigan Reading Association, 1985). The *Teacher Handbook for English Language Arts* (NCDPI, 1992) states that "an effective communication skills program must be concerned with both process and content—with how students learn and what they learn" (page 5) and that "communication skills can be developed through listening to good literature because it exposes students to models of exemplary uses of language, helps develop an awareness of the power and beauty of languages, and can provide models for good writing" (page 22).

Table 3. End-of-Course Test Development Calender.

Stage in Development Process	Algebra I	Biology	Economic, Legal, & Political Systems	English I	US History
Develop Test Specifications	May 1992	August 1993	May 1994	September 1993	August 1993
Review Items in Old Pool	Summer 1992	Aug 93 — Sept 93	June 94 — Aug 94		September 1993
Train Item Writers and Write Items	Aug 92 — Sept 92	Dec 93	July 1994	Sept 93 — Jan 94	Oct 93 — Nov 93
Edit and Review Items Items	Sept 92 — Nov 92	Dec 93 — Jan 94	Sept 94 — Oct 94	Sept 93 — Jan 94	November 1993
Item Review by Teachers Review of Item Review	December 1992		December 1994		
Edit Items/Develop New Items	January 1993	Dec 93 — Jan 94	January 1995	Jan 94 — Feb 94	Dec 93 — Jan 94
Prepare Field Tests	February 1993	February 1994	February 1995	Jan 94 — Feb 94	February 1994
Field Test Items and Teacher Review of Items	May 1993	May 1994	May 1995	May 1994	May 1994
Analyze Field Test Data	July 93 — Aug 93	July 94 — Aug 94	July 95 — Aug 95	July 94 — Aug 94	July 94 — Aug 94
Review Field Test Results and Comments	Sept 93 — Oct 93	Sept 94 — Oct 94	March 1996	Sept 94 — Oct 94	Sept 94 — Oct 94
Develop Statewide Test Forms	November 1993	November 1994	April 1996	October 1994	November 1994
Test Review by Teachers Review Test Review	January 1994	January 1995	May 1996	December 1994	December 1994
Check Answer Keys, Prepare Scale Score Tables and Manuals	Feb 94 — Mar 94	Feb 95 — Mar 95	May 96 — June 96	Feb 95 — Mar 95	Feb 95 — Mar 95
Administer Statewide Forms	May 1994	May 1995	July 1996	May 1995	May 1995

BEST COPY AVAILABLE

Reading for information allows the reader "to make personal responses and judgements about the information as part of the reading process" and students "will become more productive if [they] are taught specific strategies for collecting data or ideas, recognizing relationships, and applying information" (page 23).

Writing, as defined by the National Council of Teachers of English, is "the process of selecting, combining, arranging, and developing ideas in effective sentences, paragraphs, and, often, longer units of discorse." The stages of the writing process described in the North Carolina *Standard Course of Study* include pre-writing, drafting, revising, editing, and publishing. Editing is the stage in the process where the student prepares his or her writing for publication. In addition to examining the content and organization of the selection, the student must "locate and correct errors in punctuation, capitalization, spelling, usage, and sentence structure so that errors in conventions do not interfere with a reader's ability to understand the message" (*Teacher Handbook for English Language Arts*, NCDPI, 1992, pages 28 through 34).

Science The North Carolina Science curriculum is aligned with the national science standards developed by the National Research Council (document finalized and published in 1996). "Scienctific literacy implies an understanding of basic science concepts and the scientific processes of reasoning. The scientifically literate person has a substantial knowledge of concepts, conceptual networks, and process skills which enable the individual to continue to learn and think logically" (NCDPI, Science *Standard Course of Study*, 1994, page 4). The science curriculum (K-12) is built on the following program goals:

- Understand the nature of science.
- Become proficient is using science process skills to solve problems and make decisions.
- Develop skills to manipulate and/or operate science equipment.
- Develop responsible attitudes toward the environment, science, technology, and society.
- Understand basic scientific concepts and principles.

The science content program goal is designed to be integrated with the other four program goals; none of the program goals should be taught in isolation.

Social Studies The North Carolina Social Studies curriculum "prepares students to be active, informed, and responsible citizens of the state and the nation" (NCDPI, Social Studies *Standard Course of Study*, 1991, page 3). After completing a social studies program, students should possess the following skills:

- the understanding and acceptance of the responsibilities of citizens of a demoncratic society.
- the acquisition and use of information for problem-solving, decision-making, and planning; for self-mangagement and social participation; and for civic participation.
- the ability to apply concepts, generalizations, and theories to analyze and explain the following: (1) economic, social, and political institutions; (2) development and characteristics of societies; (3) issues and problems; and (4) geographic concepts.
- the demonstration of values consistent with the fundamental tenets of democracy.
- the exhibition of constructive attitudes toward change, conflict, diversity, and uncertainity.
- the demonstration of concern for others and the environment.

These six skills are consistent with social studies frameworks used by national indicators of social studies such as the National Assessment of Educational Progress (NAEP). The study of history places human beings and their activities in time. Knowledge of economics brings an understanding of how individuals and societies decide (and have decided in the past) to use their skills and resources to provide for their needs. Finally, knowledge of political science includes understanding political institutions: why they exist, how they function, and how each institution relates to all others.

Appendix B contains the specific goals and objectives approved by the State Board of Education for Algebra I (Mathematics curriculum adopted in 1992), English I (English Language Arts curriculum adopted in 1992), Biology (Science curriculum adopted in 1994), and ELP and US History (Social Studies curriculum adopted in 1991) as the basis for instruction.

Test Specifications

The content validity of the item pools was defined through a number of operations. First, the specifications for the Algebra I, Biology, ELP, English I, and US History item pools were defined. Working with groups of educators—NCDPI curriculum specialists, teachers, administrators, university professors, NCDPI testing consultants, and others—test specifications were established for each of the content areas assessed. The definition and refinement of the content specifications for the tests were continual processes.

Achievement test items can be classified along several dimensions. Two dimensions used to classify items for the end-of-course tests are *difficulty level* and *thinking skill level.*

Difficulty level describes how hard the item is. Easy items are ones that about 70% of the students would answer correctly. Medium items are ones that about 50% to 60% of the students would answer correctly. Finally, hard items are ones that only about 20% or 30% of the students would answer correctly.

The other classification dimension, thinking skill level, describes the cognitive skills that a student must employ to solve the problem. One item may ask a student to classify several passages based on their genre (thinking skill: organizing); another question may ask the students to select the best procedure to use for solving a problem (thinking skill: evaluating).

In order to classify items by the thinking skill required, a framework to describe thinking skills must be used. The thinking skills framework used with the end-of-course tests is from *Dimensions of Thinking* by Robert J. Marzano and others (1988). Many similar frameworks exist (for instance, that of Bloom), but *Dimensions of Thinking* was adopted by the North Carolina Department of Public Instruction in framing the revised *Standard Course of Study*. *Dimensions of Thinking* was developed through a collaborative process involving leading national experts in "thinking skills." The framework reflects current thinking in cognitive psychology, education, and philosophy. It provides a practical framework for curriculum development, instruction, assessment, and staff development.

A visual representation of the framework and a brief description of each of the dimensions of thinking are presented on the following pages. The framework should be a useful reference for curriculum development, instructional design, and in-service training.

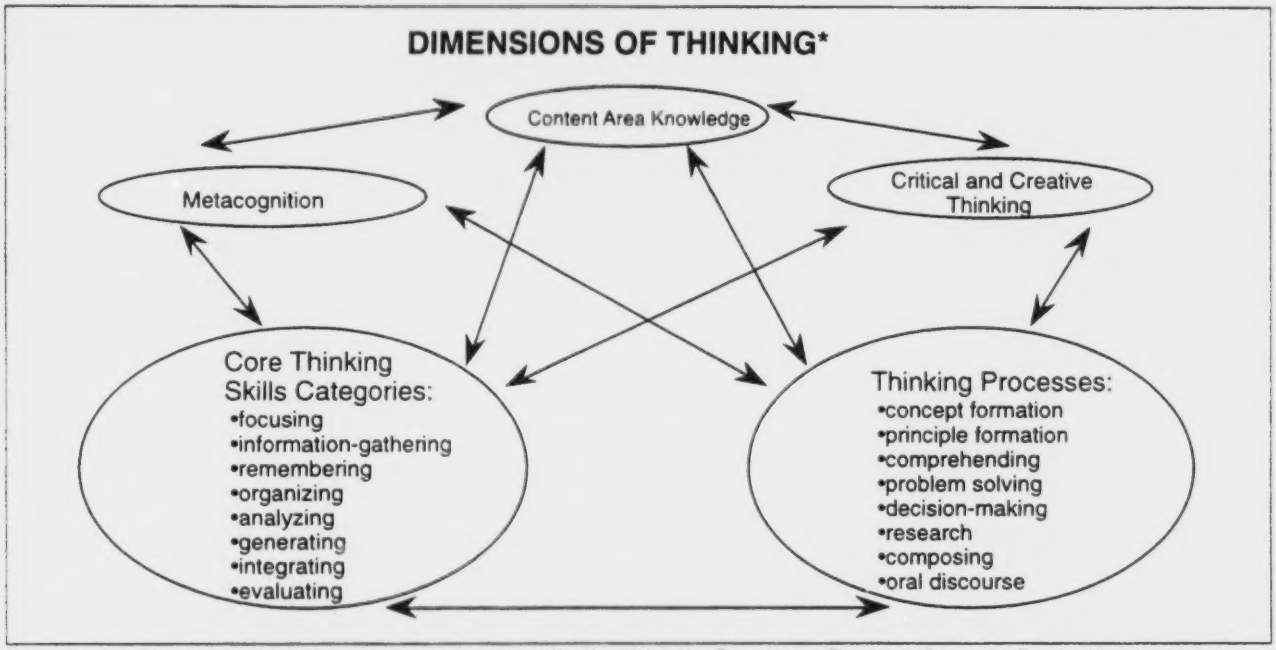


Figure 2. Thinking skills framework used with the North Carolina End-of-Course Tests (*adapted from Robert Marzano et al., *Dimensions of Thinking*, 1988).

Metacognition Metacognition refers to awareness and control of one's thinking, including commitment, attitudes, and attention.

Critical and Creative Thinking The terms "critical" and "creative" thinking are ways of describing how we go about thinking. The two are not opposite ends of a single continuum—rather, they are complementary.

1. *Critical thinking* is "reasonable, reflective, thinking that is focused on deciding what to believe or do." Critical thinkers try to be aware of their own biases, and try to be objective and logical.
2. *Creative thinking* is "the ability to form new combinations of ideas to fulfill a need" or to get "original and otherwise appropriate results by the criteria of the domain in question."

Thinking Processes A thinking process is a relatively complex sequence of thinking skills.

1. *Concept formation*: organizing information about an entity and associating that information with a label (word).
2. *Principle formation*: recognizing relationships between or among concepts.
3. *Comprehending*: generating meaning or understanding by relating new information to prior knowledge.
4. *Problem solving*: analyzing and resolving a perplexing or difficult situation.
5. *Decision-making*: selecting from alternatives.
6. *Research*: conducting scientific inquiry.
7. *Composing*: developing a product which may be written, musical, mechanical, or artistic.
8. *Oral discourse*: talking with other people.

Core Thinking Skills A thinking skill is a relatively specific cognitive operation that can be considered a "building block" of thinking. Items are classified by the following skills because they: (1) have a sound basis in research and theoretical literature, (2) are important for students to be able to do, and (3) can be taught and reinforced in school.

Knowledge (1)

Focusing Skills—attending to selected pieces of information and ignoring others.

1. *Defining problems*: clarifying needs, discrepancies, or puzzling situations.
2. *Setting goals*: establishing direction and purpose.

Information-Gathering Skills—bringing to consciousness the relevant data needed.

3. *Observing*: obtaining information through one or more senses.
4. *Formulating questions*: seeking new information through inquiry.

Remembering Skills—storing and retrieving information.

5. *Encoding*: storing information in long-term memory.
6. *Recalling*: retrieving information from long-term memory.

Organizing (4)—arranging information so it can be used effectively.

7. *Comparing*: noting similarities and differences between or among entities.
8. *Classifying*: grouping and labeling entities on the basis of their attributes.
9. *Ordering*: sequencing entities according to a given criteria.
10. *Representing*: changing the form but not the substance of information.

Applying (5)—demonstrating prior knowledge within a new situation. The task is to bring together the appropriate information, generalizations or principles that are required to solve a problem.

Analyzing (6)—clarifying existing information by examining parts and relationships.

11. *Identifying attributes and components*: determining characteristics or parts of something.
12. *Identifying relationships and patterns*: recognizing ways in which elements are related.
13. *Identifying main idea*: identifying the central element; for example, the hierarchy of key ideas in a message or line of reasoning.
14. *Identifying errors*: recognizing logical fallacies and other mistakes and, where possible, correcting them.

Generating (7)—producing new information, meaning, or ideas.

15. *Inferring*: going beyond available information to identify what reasonably may be true.
16. *Predicting*: anticipating next events, or the outcome of a situation.
17. *Elaborating*: explaining by adding details, examples, or other relevant information.

Integrating (8)—connecting and combining information.

18. *Summarizing*: combining information efficiently into a cohesive statement.
19. *Restructuring*: changing existing knowledge structures to incorporate new information.

Evaluating (9)—assessing the reasonableness and quality of ideas.

20. *Establishing criteria*: setting standards for making judgements.
21. *Verifying*: confirming the accuracy of claims.

Algebra I The items for the Algebra I item pool were specified by goal and objective, difficulty level, and thinking skill. Table 4 shows the content specifications for the Algebra I test. Within goals, objectives were weighted equally in the test specifications. This is consistent with traditional aspects of curriculum development: the more important a goal is considered to be, the greater the number of objectives that will be developed for it.

For difficulty level, 25% of the items were specified to be written at the easy level, 50% of the items were specified to be written at the medium level, and 25% of the items were specified to be written at the difficult level. The thinking skill level for each item was associated with the content objective. For example, the sample item in Appendix D was written for objective 9.10, "Determine if a set of data respresents an exponential function," and the thinking skill was specified as "Analyzing." With the greater emphasis on solving problems "in context" and using "real-world" applications, the test requires more reading than typically found on standardized achievement tests.

The objectives assessed on the calulator part of the test were identifed by the NCDPI curriculum specialists (see Appendix B concerning which objectives are calculator-active indicated by a "c" next to the objective number). This part of the test assesses a student's ability to solve problems rather than apply specific procedures. Students are expected to have a graphing calculator for use during instruction and on the test. Students are not directed to use the calculator to solve a problem; students must decide if and when to use it.

Table 4. Item pool specifications for the North Carolina End-of-Course Test of Algebra I—Percent of items on the test assessing each goal of the curriculum.

Goal	Description of Goal	Percent of Items in Pool and on Test
1	The learner will use the language of algebra.	5%
2	The learner will perform operations with real numbers.	12%
3	The learner will solve equations and inequalities with one variable.	16%
4	The learner will demonstrate an elementary understanding of relations and functions.	12%
5	The learner will graph and use linear equations and inequalities.	14%
6	The learner will graph and solve systems of linear equations and inequalities.	9%
7	The learner will perform operations with polynomials.	14%
8	The learner will work with ratios, proportions, and percents.	5%
9	The learner will explore, graph, and interpret nonlinear equations.	14%

The first version of the Algebra I test, administered from 1986 to 1993, was aligned with the 1985 *Standard Course of Study* for Algebra I. That version of the test focused more on specific procedures rather than problem solving. The 1992 revison of the *Standard Course of Study* moved many of the elementary algebra concepts (e.g., solving simple linear equations) back to the algebra strand (goal 3) of the mathematics curriculum for grades 7 and 8.

Biology Each item for the Biology item pool was specified by content goal and objective, skill goal, difficulty level, and thinking skill.

Each of the content goals and objectives were not weighted equally in the test specifications. Each content goal and objective was examined by NCDPI curriculum specialists and weighted appropriately. The content goals were specified as follows:

- Goal 6—Chemical and Physical Basis of Life (30% of the item pool);
- Goal 7—Continuity of Life (40% of the item pool); and
- Goal 8—Ecology (30% of the item pool).

See Appendix B for the weighting of each of the content objectives.

In addition to the content of each item being specified, the skill employed to answer each item was also specified. The skill goals are defined as follows:

- Goal 1—*Nature of Science* (5% of the item pool). An understanding of the nature of science includes an appreciation of the methods of scientific investigation. A scientifically literate person understands the following about the nature of science: science is public, science is historic, science is replicable, science is tentative, and science is probabilistic.
- Goal 2—*Process Skills* (at least 50% of the item pool). Science process skills provide a foundation for scientific inquiry, supporting the development of reasoning and problem-solving abilities. Process skills are usually acquired through active experiential learning. The following process skills are assessed: observing, classifying, using numbers, communicating, measuring, inferring, predicting, using space/time relations, interpreting data, defining operationlly, experimenting, controlling variables, formulating hypotheses, and formulating models.
- Goal 3—*Manipulative Skills* (5% of the item pool). It is important that students know how to use laboratory equipment safely. They should also know how to choose equipment suitable for the experiment they have designed, and how to assemble and care for that equipment. Students also need to know how to properly handle and care for living organisims.
- Goal 5—*Current Topics in Science* (5% of the item pool). Knowledge of current environmental and societal issues lends depth to the student's understanding of the curriuclum. It is important that the student be taught about current research, which will relate not only to the curriculuar material but also to career potential in research or applied science.
- *Conceptual Knowledge* (at most 35% of the item pool). Conceptual knowledge reflects knowledge of the fundamental facts, concepts, and theories of science.

For difficulty level, 25% of the items were specified to be written at the easy level, 50% of the items were specified to be written at the medium level, and 25% of the items were specified to be written at the difficult level. The thinking skill level for each item was associated with the content objective. For example, item #4106 in Appendix A was written for objective 6.3, "Demonstrate knowledge of the chemical processes of life," and the thinking skill was specified as "Evaluating."

Goal 4 of the Biology curriculum states that the learner should develop a responsible attitude toward the environment, science, technology, and society. This goal is not assessed by the Biology end-of-course test. Periodically, students and teachers will be surveyed concerning their attitudes and behaviors.

The original version of the Biology test was administered from 1987 to 1994 and was aligned with the 1985 *Standard Course of Study* for Biology. That version of the test focused on the knowledge of facts and ideas, whereas this version examines whether a student can go beyond memorization and actually apply process skills to the investigation of science.

Economic, Legal, and Political Systems The items for the ELP item pool were specified by goal and objective, difficulty level, and thinking skill. Table 5 shows the content specifications for the ELP test. All of the goals were weighted evenly except goal 10 which had half as much weight as the other nine goals. Within goals, the objectives were weighted equally in the test specifications.

For difficulty level, 25% of the items were specified to be written at the easy level, 50% of the items were specified to be written at the medium level, and 25% of the items were specified to be written at the difficult level. The thinking skill level for each item was associated with the content objective. For example, item #R008 in Appendix A was written for objective 4.4, "Demonstrate how supply, demand, and competition affect prices and the availability of goods and services," and the thinking skill was specified as "Organizing."

Table 5. Item pool specifications for the North Carolina End-of-Course Test of ELP—
Percent of items on the test assessing each goal of the curriculum.

Goal	Description of Goal	Percent of Items in Pool and on Test
1	The learner will investigate issues and problems confronting the American economic, legal and political systems.	10%
2	The learner will explain rights and analyze the obligations of responsible citizenship.	11%
3	The learner will investigate how and why individuals and groups make economic choices.	10%
4	The learner will analyze features of the economic system of the United States.	11%
5	The learner will analyze factors influencing the United States economy.	11%
6	The learner will explain the function and importance of the North Carolina and United States Constitutions.	11%
7	The learner will explain the structure and functions of local, state, and national governments and understand their relationship.	10%
8	The learner will explain why laws are needed and how they are enacted, implemented, and enforced.	11%
9	The learner will explain how the political and legal systems provide for balancing competing interests and resolving conflicts.	10%
10	The learner will evaluate the influence of ethical and moral principles and religious beliefs on the development of our economic, legal, and political systems.	5%

The first version of the ELP test was administerd from 1990 to 1996 and was based on the 1985 *Standard Course of Study*. That version of the test focused on the knowledge of facts and ideas, whereas this version examines whether a student can go beyond the memorization of people, events, and facts to apply the knowledge and skills they have learned to other events.

English I All of the items for the English I test are passage-dependent; both editing and textual analysis are assessed in context.

The editing part of the test assesses one aspect of the writing strand of the English Language Arts curriculum—the extent to which students can determine if a word or phrase is incorrect and, if so, what should the correction be. The editing item pool consists of passages (released student essays from the English II test) which are followed by 5 to 8 questions. Words or phrases are underlined and numbered in each passage and the associated questions consist of three possible corrections or the option "Make no change." The errors are specified according to the English II Conventions Scoring Criteria in Table 6.

Table 6. English II Conventions Scoring Criteria as applied to the editing item pool of English I.

Description of Criteria	Percent of Items in Pool
Sentence Formation, including fragments, run-ons, misplaced or dangling modifiers, incorrect subordination, and non-parallel structure.	40%
Usage, including subject/verb agreement, tense, verb inflections, pronoun antecedent agreement, case of pronouns, pronoun reference, apostrophes.	30%
Mechanics, including capitalization, end punctuation, internal punctuation, and paragraphs.	20%
Spelling.	10%

The textual analysis part of the test assesses primarily goals 2 and 3 of the the reading and viewing strands of the English Language Arts curriculum. Students must apply their knowledge of literary terms and elements to a variety of passages. The passages were selected based on several criteria: they must be interesting to read, be complete (with a beginning, middle, and end), and be from sources students might actually read. By adhering to these criteria in the selection of passages, the passages tend to be longer than those typically found on tests. The passages for each form on this part of the test were selected to represent a theme (e.g., freedom, making choices, change). A thematic approach is helpful in getting students to see the relationships between things and to obtain deeper thinking about a topic. However, adherence to a thematic approach was not so strict as to limit the ability to cover the curriculum or lead to items being forced into a theme when it seems artificial.

The items for the textual analysis part of the item pools were specified to be appropriate to the passage so that the relevant aspects of the passage were assessed. Specific goals and objectives were not specified in advance for each passage according to prepared test specifications, but most passages should have items spanning the goals and objectives of the English Language Arts curriculum. The passages have some reading comprehension

items, but the majority of the items require the application of literary terms and elements. The terms and elements involve a knowledge of genres, such as epic poetry, tragedy, comedy, short stories, novels, editorials, and newspaper articles; and the application of appropriate terms and elements to analyze the different genres. For example, in a study of tragedies, such terms as tragic hero and tragic flaw would be appropriate. In a newspaper editorial, an author's tone and point of view, presence or absence of bias, and use of propoganda techniques might be analyzed. In short stories, narrative techniques like flashback, foreshadowing, and dramatic irony would be included. Goal 4, personal response, is not assessed by the English I multiple choice test. This goal is better assessed in an open-ended format.

The previous version of the English I test (administered from 1990 to 1994) assessed reading comprehension, grammar and editing, and identification and application of literary terms and elements. This version of the test assesses the same domains, but in a different format. Application of literary terms and elements occurs in the analysis of text, rather than in isolation. Similarly, editing involves revising student essays rather than just the recognition of errors.

US History The items for the US History item pool were specified by goal and objective, difficulty level, and thinking skill. Table 7 shows the content specifications for the US History test. The items in the pool are distributed as follows: 40% of the items assess knowledge of the foundation of the United States to the Civil War and Reconstruction (goals 1 through 5) and 60% of the items assess knowledge of post-Reconstruction to the present (goals 6 through 12). Within goals, objectives were not weighted equally in the test specifications. Each objective was examined by the NCDPI curriculum specialists and weighted appropriately (see Appendix B for the weighting of individual objectives).

For difficulty level, 25% of the items were specified to be written at the easy level, 50% of the items were specified to be written at the medium level, and 25% of the items were specified to be written at the difficult level. The thinking skill level for each item was associated with the content objective. For example, item #2882 in Appendix A was written for objective 6.4, "Evaluate the effects of racial segregation on various regions and segments of American society," and the thinking skill was specified as "Generating."

The first version of the US History test was administerd from 1988 to 1994 and was based on the 1985 *Standard Course of Study*. That version of the test focused on the knowledge of facts and ideas, whereas this version examines whether a student can go beyond the memorization of people, events, and facts to apply the knowledge and skills they have learned to other events. While knowledge of facts and concepts is important, the items on the revised test are at a much broader level and concern major ideas that students are expected to know to be considered literate.

Table 7. Item pool specifications for the North Carolina End-of-Course Test of US History—Percent of items on the test assessing each goal of the curriculum.

Goal	Description of Goal	Percent of Items in Pool and on Test
1	The learner will analyze those elements in the American colonial experience that led to separation from England.	6%
2	The learner will apply ideas of self-government as expressed in America's founding documents.	9%
3	The learner will judge the effectiveness of the institutions of the new nation in completing its independence (1781-1815).	9%
4	The learner will assess the competing forces of nationalism and sectionalism in the periods 1815-1850.	9%
5	The learner will evaluate the Civil War and Reconstruction as an affirmation of the power of the national government.	6%
6	The learner will interpret economic, social and political trends of the late 19th and early 20th centuries.	9%
7	The learner will analyze the causes and effects of US involvement in international affairs.	8%
8	The learner will appraise the economic, social, and political changes of the decades of the '20s and '30s.	9%
9	The learner will analyze and evaluate the significance of causes, events, and effects of the WW II Era.	8%
10	The learner will trace economic and social developments and assess their significance for the lives of Americans in the last half century.	9%
11	The learner will analyze changes in American political life in the last half century.	9%
12	The learner will evaluate the conduct of United States foreign policy over the last half century.	9%

Item Writing and Review

Review of Old Item Pools Prior to item writing, the first task in the process was to examine the old item pools for possible items that were aligned with the revised *Standard Course of Study* for Algebra I, Biology, ELP, and US History. Each item was reviewed by NCDPI curriculum specilaists and other curriculum specialists. Table 8 shows the number of items that were retained for each of the four item pools.

Selection and Training of Item Writers Algebra I, Biology, ELP, and US History item writers were nominated by the NCDPI curriculum specialists from across the state because of their knowledge of the curriculum and exemplary teaching status. Fifteen teachers and curriculum specialists from each content area were trained in the technical aspects of item writing. Each item writer was sent a training packet specially designed for the subject matter for which they were to write items. The materials consisted of a videotape (and a script) of *How to Write Multiple Choice Achievement Test Items* developed by the NCDPI, three packets of work materials related to the specific content area (i.e., Algebra I), and a copy of *Guidelines for Bias-Free Publishing* developed by CTB/ McGraw-Hill. As described by Haladyna (1994), the "best-answer" format of multiple choice questions was also used in these tests. This format is well suited for testing a student's ability to evaluate (Marzano's highest thinking skill level).

Each item writer was asked to submit 10 items as part of the training phase. These 10 items were evaluated and, based on the evaluations by content specialists, the item writers were asked to develop additional items. Item writers received feedback on the 10 items they had developed to help them better develop the additional items requested. Because of the refocusing on higher order thinking skills and "real world" applications, the Biology, ELP, and US History item writers met as a group to discuss the revisions in the curriculum and the types of items they should write, and to review items previously developed.

For English I, an Advisory Group was establised that consisted of NCDPI Testing and Instructional consultants, curriculum specialists from local school systems, and classroom teachers. Of the ten members, more than half were currently teaching English I. Due to the thematic structure of the English I test and item pool, one curriculum specialist was contracted to develop the items for the textual analysis part of the test. The consultant worked with the Advisory Group to determine themes appropriate for grade 9 students, to select and review passages, to suggest ideas for items, and then to review the items developed by the consultant. In addition, because of the structure of the test and item pool, the grammar items were also developed by one curriculum specialist. The Advisory Group specified the types of questions to be developed, the weighting of the English II Conventions Scoring Criteria for the English I item pool, and reviewed and edited all of the items developed by the consultant.

The use of classroom teachers from across the state as item writers and developers helped to insure that instructional validity was maintained because the background would be drawn from their classroom experiences.

Selection of Passages for English I Passages for the textual analysis part of the test were selected by the English I Advisory Group. The selected passages were ones that were related to a theme appropriate for grade 9 students and were ones that would generally be read by students, would be interesting to students, and were appropriate content for an achievement test.

For each passage, a frame was written as an introduction to the passage. Each frame was designed to stir interest and support comprehension, while at the same time not reveal something in the passage that the reader should discover for himself or herself.

The evaluation of each passage and the associated questions included the following:

- Is the passage interesting to read and does it have a beginning, middle, and end?
- Is the frame acceptable for the passage?
- Do all of the objectives fit well with the passage, or should one or more not be used and substituted with another objective?
- Do the items adequately cover the major content of the passage? Are the most important ideas included?

Item Writing and Review Each item writer was contracted to write 75 items from across the curriculum according to the specifications on each Item Specifications Form. Table 8 shows the number of new items developed for each subject area. See Appendix D for a completed Item Specification Form for a sample item.

Table 8. Number of items developed by subject area.

Subject Area	Items Retained from Old Item Pools	New Items Developed	Passages Selected
Algebra I	685	760	—
Biology	155	907	—
ELP	630	437	—
English I	—	850	88
US History	458	1,042	—

Next, the item pools were analyzed by curriculum specialists and classroom teachers to assure that the items were valid representations of the objectives for which they were written. For each subject area, 10 to 15 individuals reviewed the items (and passages for English I). Item reviewers learned about the revised *Standard Course of Study* for the subject area for which they would be reviewing items, the end-of-course testing program, and the test development process in general. Each item reviewer received a copy of the document *How to Review Multiple Choice Achievement Test Items* (developed by NCDPI) which described the criteria to be used to evaluate each item.

The criteria for evaluating each item included the following:

- conceptual: objective match, fair representation, lack of cultural bias, clear statement, single problem, one best answer, common context in foils, each foil credible;
- language: appropriate for age; correct punctuation, spelling, and grammar; lack of excess words; no stem/foil clues; no negative in foils;

- format: logical order of foils; familiar presentation style, print size, and type; correct mechanics and appearance; equal length foils; and
- diagram: necessary, clean, relevant, unbiased.

Each item was reviewed by at least four individuals. See Appendix D for a sample completed Item Review Form for an Algebra I item and the summary of the teacher comments.

Next, the results from the item reviews were aggregated by item (and passage). The results were then examined item-by-item by NCDPI curriculum specialists and test development staff. Based on the comments from the reviewers, items were revised and/or rewritten, item-objective matches were examined and changed where necessary, and frames and diagrams for passages were refined.

Additional passages were found and additional items were developed where necessary. These items were reviewed by curriculum specialists and test development staff based on the same criteria as the first review.

Field Testing

The next stage of the item development process was to collect the items for each subject area into test forms for field testing. Although the forms were not the final forms of the North Carolina End-of-Course Tests, they were organized according to the specifications for the final tests as shown in Table 4 for the Algebra I test, the discussion on pages 18 and 19 for the Biology test, Table 5 for the ELP test, Table 6 and the discussion on pages 20 and 21 for the English I test, and Table 7 for the US History test. Table 9 shows the number of field test forms and the number of items field tested for each subject area.

Table 9. Number of items and passages field tested in each subject area.

Subject Area	Number of Field Test Forms	Number of Items Field Tested	Number of Passages Field Tested
Algebra I	14	1,259	—
Biology	11	1,062	—
ELP	11	957	—
English I	11	777	80
US History	14	1,470	—

Next, test administration instructions were written, distribution procedures were organized, and administrators were trained to conduct the field test administration. The test administration organization used to administer statewide tests in North Carolina was employed to do the field testing. The administration of the field test forms followed the routine eventually expected to be used when the statewide tests were administered.

Samples of students were selected to take the field test forms. To insure broad representation, schools were selected from across the state and were representative of the state based on the ethnic/racial characteristics of the student population, geographic location, and scores on previous versions of the tests. Table 10 shows the characteristics of the field test samples for Algebra I, Biology, ELP, English I, and US History (number of students tested, number of schools sampled, and percent of students in the field test samples that were idendified as limited English proficient). While every effort was made to ensure full participation, only limited modifications for exceptional children were available for use during the field tests (e.g., dictation to a proctor/scribe, magnification devices, student marks in the test book, multiple test sessions, scheduled extended time, and testing in a separate room). Modifications not availiable during field testing included braille and large-print versions of the test books.

Table 10. Characteristics of the field test samples.

Subject Area	Number of Students Tested	Number of Schools Sampled	Percent of Students Tested Identified as LEP
Algebra I	10,103	130	0.5
Biology	9,742	63	0.6
ELP	9,664	57	0.7
English I	7,334	49	0.5
US History	10,106	63	1.4

Item Analysis and Selection

The field test data for all items were analyzed by the NCDPI using both the classical measurement model and the three-parameter logistic item response theory model. Item statistics and descriptive information (item number, passage number, field test form and item number, curriculum objective, and answer key) were printed on labels and attached to the item record for each item. The item record contains the statistical, descriptive, and historical information for an item; a copy of the item itself as it was field tested; any comments by reviewers; and the psychometric notations. Each item has a separate item record. See Appendix D for the item record from of a sample item.

Classical Measurement Analyses For each item the p-value (percent correct), the standard deviation of the p-value, and the point-biserial correlation between the item score and the total test score were computed using SAS. In addition, frequency distributions of the response choices were tabulated.

Item Response Theory Analyses Classical test theory has two basic shortcomings: (1) the use of item indices whose values depend on the particular group of examinees from which they were obtained, and (2) the use of examinee ability estimates that depend on the particular choice of items selected for a test. The basic premises of item response theory (IRT) overcome these shortcomings by predicting the performance of an examinee on a test item based on a set of underlying abilities. The relationship between an examinee's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases.

The three-parameter logistic model (3PL) takes into account the difficulty of the item and the ability of the examinee. An examinee's probability of answering a given item correctly depends on the examinee's ability and the characteristics of the item. The one-parameter model (Rasch) only takes into account the difficulty of the item. The 3PL model has three assumptions: (1) unidimensionality—only one ability is assessed by the set of items, (2) local independence—when abilities influencing test performance are held constant, an examinee's responses to any pair of items are statistically independent (conditional independence, i.e., the only reason an examinee scores similarly on several items is because of his or her ability, not because the items are correlated), and (3) the item characteristic curve (ICC) specified reflects the true relationship among the unobservable variable (ability) and the observable variable (item response). The equation for the three-parameter logistic model is

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad \text{where } i = 1, 2, ..., n \qquad \text{(Equation 1)}$$

$P_i(\theta)$—is the probability that a randomly chosen examinee with ability θ answers item i correctly (this is an S-shaped curve with values between 0 and 1 over the ability scale)

a—the slope or the discrimination power of the item (the slope of a typical item is 1.00)

b—the threshold or the point on the ability scale where the probability of a correct response is 50% (the threshold of a typical item is 0.00)

c—the asymptote or the proportion of the examinees who got the item correct, but did poorly on the overall test (the asymptote of a typical 4-choice item is 0.20)

D—a scaling factor to make the logistic function as close as possible to the normal ogive function (equals 1.7)

The IRT parameter estimates for each item were computed using the BIMAIN computer program (Muraki, Mislevy, & Bock, 1991) using the default Bayesian prior distributions for the item parameters [a~lognormal(0, 0.5), b~N(0,2), and c~Beta(6,16)].

The following figures show the item characteristic curves for a typical 4-option multiple-choice item and several items from the end-of-course item pools.

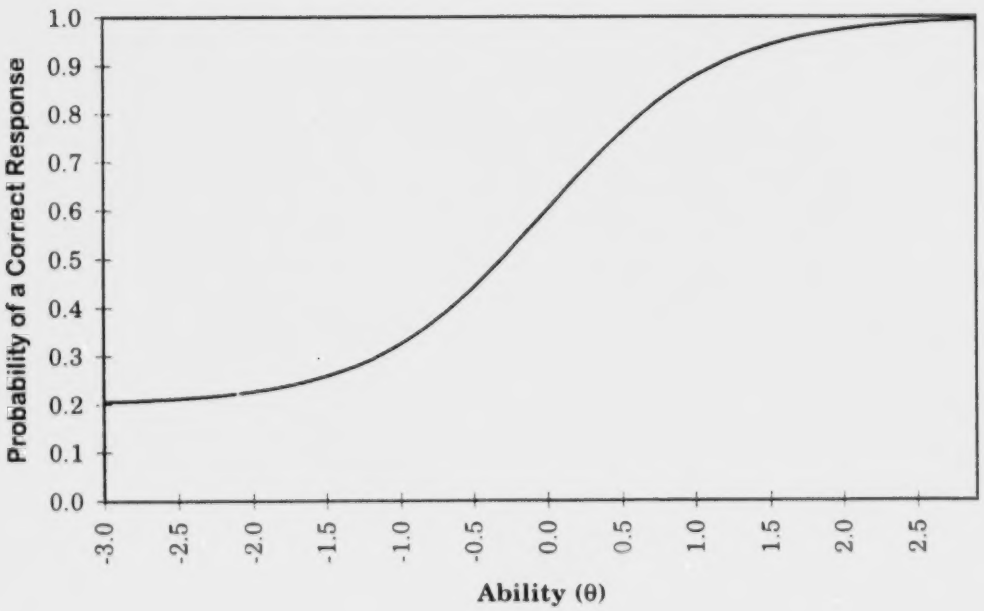


Figure 3. Item characteristic curve of a typical 4-option multiple-choice item (a = 1.00, b = 0.00, and c = 0.20).

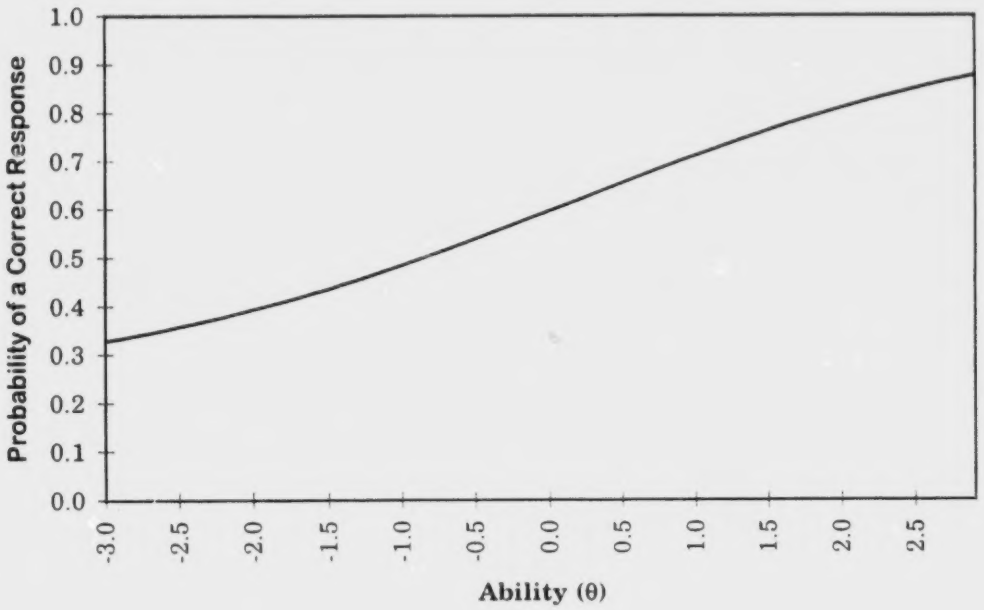


Figure 4. Item characteristic curve of an item that exhibited a low slope (Algebra I item #2098, $a = 0.357$, $b = 0.168$, and $c = 0.229$). This item was flagged as exhibiting "Weak Prediction."

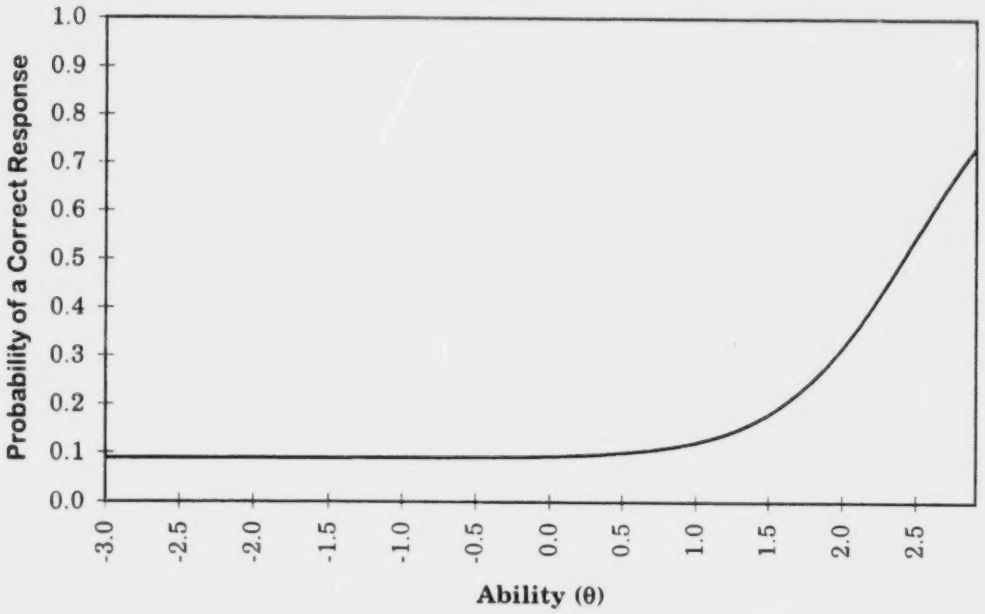


Figure 5. Item characteristic curve of an item that was difficult, but was retained for test development (Biology item #4106, $a = 1.298$, $b = 2.499$, and $c = 0.089$).

Bias Analyses Differential item functioning (DIF) examines the relationship between the score on an item and group membership while controlling for ability. The Mantel-Haenszel procedure examines DIF by examining j 2×2 contingency tables, where j is the number of different levels of ability actually achieved by the examinees (actual total scores received on the test). The focal group is the focus of interest and the reference group serves as a basis for comparison for the focal group (Dorans and Holland, 1993; Camilli and Shepherd, 1994).

The Mantel-Haenszel chi-square statistic tests the alternative hypothesis that there is a linear association between the row variable (score on the item) and the column variable (group membership). The χ^2 distribution has 1 degree of freedom and is determined as

$$Q_{MH} = (n-1)r^2 \qquad \text{(Equation 2)}$$

where r^2 is the Pearson correlation between the row variable and the column variable (SAS Institute, 1985).

The Mantel-Haenszel (MH) Log Odds Ratio statistic is used to determine the direction of differential item functioning (DIF) in SAS. This measure is obtained by combining the odds ratios, α_j, across levels with the formula for weighted averages (Camilli and Shepherd, 1994, p. 110):

$$\alpha_j = \frac{p_{Rj} / q_{Rj}}{p_{Fj} / q_{Fj}} = \frac{\Omega_{Rj}}{\Omega_{Fj}} \qquad \text{(Equation 3)}$$

For this statistic, the null hypothesis of no relationship between score and group membership, or that the odds of getting the item correct are equal for the two groups, is not rejected when the odds ratio equals 1. For odds ratios greater than 1, the interpretation is that an individual at score level j of the Reference Group has a greater chance of answering the item correctly than an individual at score level j of the Focal Group. Conversely, for odds ratios less than 1, the interpretation is that an individual at score level j of the Focal Group has a greater chance of answering the item correctly than an individual at score level j of the Reference Group. The Breslow-Day Test is used to test whether the odds ratios from the j levels of the score are all equal. When the null hypothesis is true, the statistic is distributed approximately as a χ^2 with j-1 degrees of freedom (SAS Institute, 1985).

For the end-of-course tests, males (approximately 49.4% of the population) and blacks (approximately 28.6% of the population) were defined as the focal groups and females (approximately 50.5% of the population) and whites (approximately 65.4% of the population) were defined as the reference groups.

Criteria for Inclusion in Item Pools Items were flagged as exhibiting psychometric problems or bias due to ethnicity/race or gender according to the following criteria:

- "weak prediction"—the slope (a parameter) was less than 0.60,
- "guessing"—the asymptote (c parameter) was greater than 0.40,
- "ethnic" bias—the log odds ratio was greater than 1.5 (favored whites) or less than 0.67 (favored blacks), and
- "gender" bias—the log odds ratio was greater than 1.5 (favored females) or less than 0.67 (favored males).

The ethnic and gender bias flags were determined by examining the significance levels of items from several forms and identifying a typical point on the continuum of odds ratios that was statistically significant at the $\alpha = 0.05$ level. Because the tests were to be used to evaluate the implementation of the curriculum, items were

not flagged on the basis of the difficulty of the item (threshold).

Review Of Item Pools During the field test each test administrator was asked to review one field test form (the top one in the set of materials they received for administration). The teachers were asked to respond to two questions and to make any general comments concerning each item on the form.

1. Instruction: Which of the following describes the concept or skill measured by this item?
   - **Basic** to instruction this year and taught to most students in this class. When coding an item as "basic," consider whether classroom instruction has been sufficient such that the item is a fair test question for most students in the class;
   - **Enrichment** material taught only to advanced students this year; or
   - New part of the curriculum this year so **not taught**.

2. Item Quality: In your opinion, is this item appropriate for the end-of-course test? In judging whether an item is appropriate, the reviewer should take into consideration conceptual quality, language quality, format and graphics quality, and cultural bias. For any item that you believe needs revision or is not appropriate, please use the space next to the bubble to write a comment or suggestion for improvement.
   - **Yes**, if the item is okay;
   - **Revise**, if the item needs additional work; or
   - **No**, if the item should not be used in any form.

All items, statistics, and comments were reviewed by curriculum specialists and testing consultants, and items that were not deemed appropriate for curricular or psychometric reasons were deleted.

Items flagged for exhibiting ethnic and/or gender bias (Mantel-Haenszel indices greater than 1.5 or less than 0.67) were then reviewed by a group of individuals (not the developers) that represented various minority groups. The individuals on the bias review team were selected because of their minority group membership or their experience with exceptional students and because of their knowledge of the curriculum area. The members of the team were provided copies of the item records of items flagged as being biased (each item record included a copy of the item, the curricular objective the item was assessing, and the various item statistics from the field test). The team members were asked to review individually each item in terms of the following questions:

- Does the item contain any offensive gender, ethnic, and/or regional content?
- Does the item contain gender, ethnic, or cultural stereotyping?
- Does the item contain activities that will be more familiar to one group than another?
- Do the words in the item have a different meaning in one group than in another?
- Could there be group differences in performance that are *unrelated* to proficiency in the content area?

The team members were then instructed that if their answer was "yes" to any of the questions for a particular item, that they should record the 5-digit item number and check the appropriate column(s) on the Item Bias Review Sheet (see Appendix F for samples of the bias review materials). If an item was deemed to be biased, the team members were also asked to explain their decision. Items that were consistently identified as exhibiting bias were deleted from the item pool and were not used in the development of any tests. All of the biased items that were flagged, but not rejected by the bias review team, were examined by the curriculum specialists. If all students were expected to master the content of the item, then the item was retained for test development.

Final average item pool parameter estimates for each of the end-of-course tests are presented in Table 11.

Table 11. Average item pool parameter estimates for the North Carolina End-of-Course Tests.

Subject Area	IRT Parameters			P-value	Bias Odds Ratios	
	Threshold (*b*)	Slope (*a*)	Asymptote (*c*)		Ethnic/Race	Gender
Algebra I	0.87	0.97	0.21	0.46	1.01	1.04
Biology	1.13	0.88	0.21	0.44	1.03	1.02
ELP	0.63	1.01	0.21	0.50	1.07	1.02
English I	0.69	0.97	0.20	0.50	1.08	0.99
US History	1.19	0.98	0.20	0.40	1.00	1.03

Dimensionality of Item Pools The dimensionality of the end-of-course tests was examined by the L.L Thurstone Laboratory at the University of North Carolina at Chapel Hill. The analyses involved the application of weighted least squares analysis of the polychoric correlations computed among item parcels constructed from the binary items on the tests. This procedure was used to avoid the well-documented artifacts that may arise when binary data are factor-analyzed (Mislevy, 1986), and the less well-documented difficulties that arise with full-information item factor analysis (Bock, Gibbons, and Muraki, 1988). To examine the contributions of factors associated with subsets of the item parcels, the Schmid and Leiman (1957) representation of hierarchical factor analysis was used to explore the relative sizes of unique factors over and above the general factor for the test.

For all of the end-of-course tests, item parcels (Cattell, 1956) were constructed of the items measuring each curricular goal that the items were developed to measure. In all cases, item-parcel scores were computed as the summed score (number correct) for each parcel. Polychoric correlations were computed among the item-parcel scores using the computer software PRELIS (Joreskog & Sorbom, 1986), and the weighted least squares analysis was done with LISREL (Joreskog & Sorbom, 1988).

For Algebra I, generally there were two unique factors that could be identifed—one for goals 1 through 4 and the other for goals 6 through 9. For the most part, loadings for each item parcel were higher for the general factor than they were on the unique factor.

For Biology, generally there were two unique factors that could be identifed—one for goal 6 and one for goal 8. The items associated with goal 7 loaded only on the general common factor. For the most part, factor loadings for each item parcel were higher for the general factor than they were on the unique factor.

For ELP, generally two unique factors could be identifed—one for goals 1 through 5 and one for goals 6 through 10. For the most part, factor loadings for each item parcel were higher for the general factor than they were on the unique factor. For some of the field test forms used in the analysis, item parcels for goals 4, 5, and 6 were

not reliabily associated with any factor except the general factor; for other forms, one of the additional factors could not be identified.

The item parcels for English I were constructed of the items associated with each passage on the test. For several of the field test forms, a single-common-factor model fit the data very well, as measured by the likelihood-ratio goodness of fit criterion. When this model did not fit, a model with very small loadings on unique factors for item parcels 1 through 4, with a second unique factor for item parcels 5 through 9, was satisfactory.

For US History, generally only one unique factor could be identified, for goals 6 through 12, in addition to a general factor for all item parcels. For the most part, factor loadings for each item parcel were higher for the general factor than they were on the unique factor.

Because of the near-perfect covariation between the serial position of the items on the field test forms and their curricular objective classification, it is possible (or even likely) that the unique factors that could be identifed in the analyses simply reflected individual differences in time-management over the course of the tests, with some students performing better on items earlier in the test and others performing relatively better later. Alternatively, these factors may have been induced by local differneces in the timing in the academic year of, or emphasis on, the material included in specific objectives. In either case, while these factors are of data-analytic interest and point to the importance of the maintenance of balance across curricular goals in the operational forms, the existiance of more than one factor in the analyses of the field test forms does not preclude reporting a single (scaled score) on the tests.

For each end-of-course test six (nine forms for Algebra I and US History) were prepared for administration to students in each subject area. Three forms were developed for administration the first year and three forms were developed for administration during the second year. In order to assure that each set of three forms for a subject area was equivalent (both within and between the sets), the item pools were randomly split in half and the second half of each item pool was put away and not used during the development of the first set of three forms.

Items for the Algebra I, Biology, ELP, and US History tests and passages for the English I test were selected using a modified domain sampling model, with the various forms equivalent. In the modification used here, the domain of items for each test was limited to those items that had satisfactory psychometric characteristics and curricular approval and were approved by the bias review team. This was determined by the analyses of the item field-test data and the reviews by curriculum specialists and testing consultants. Some items that did not meet the psychometric criteria were used in test development because these were items that assessed new parts of the curricula (typically the items had low slopes because the material had not been completely taught to students, such as Algebra I item #2098 in Appendix A).

The three forms of each subject area test were developed according to the test specifications delineated during the initial phase of development (see the earlier discussions and Appendix B) and the average p-value for each test or subtest was equivalent to the average p-value of the entire item pool. The subtests for each test were defined by the factor analyses performed on the field tests forms or to aid in interpretation when the tests were given. The following subtests were defined and the average p-values are presented in Table 12:

- Algebra I
  (1) Basic operations, equations, and inequalities (goals 1, 2, and 3)
  (2) Functions and graphing (goals 4, 5, and 6)
  (3) Polynomials and non-linear equations (goals 7, 8, and 9)
- Biology
  (1) Chemical and physical basis of life (goal 6)
  (2) Continuity of life (goal 7)
  (3) Ecology (goal 8)
- ELP
  (1) Economics (goals 1, 2, 3, 4, and 5)
  (2) Government (Goals 6, 7, 8, 9, and 10)
- English I
  (1) Editing
  (2) Textual Analysis

Table 12. Average p-value for the total test and each subtest of the North Carolina End-of-Course Tests.

Subject Area	Total Test Average P-value	Subtest 1 Average P-value	Subtest 2 Average P-value	Subtest 3 Average P-value
Algebra I	0.456	0.514	0.443	0.407
Biology	0.425	0.460	0.413	0.401
ELP	0.499	0.522	0.474	—
English I	0.507	0.566	0.466	—
US History	0.401	—	—	—

The item parameters for each form (threshold and slope) were examined to determine if they were approximately equivalent. If the average item parameters were not equivalent, then some items were replaced with other items from the item pools to insure equivalence.

After each test was assembled into forms (3 forms for each subject area), the forms were reviewed by ten to fifteen subject area teachers and curriculum supervisors. Each group (separate by subject area) met in Raleigh for one day and worked independent of the test developers. The test review groups discussed the revised *Standard Course of Study* for the test they would be reviewing, the end-of-course testing program, and the test development process in general. During review training, each of the criteria for evaluating the tests was discussed.

The criteria for evaluating each group of three forms included the following:

- that the content of the test forms should reflect the goals and objectives of the North Carolina *Standard Course of Study* for the subject (curricular validity);
- that the content of the test forms should reflect the goals and objectives taught in North Carolina schools (instructional validity);
- that the items should be clearly and concisely written, and the vocabulary appropriate to the target age level (item quality);
- that the content of the test forms should be balanced in relation to ethnicity, gender, socioeconomic status, and geographic district of the state (test/item bias); and
- that each item should have one and only one answer that is right; however, the distractors should appear plausible for someone who has not achieved mastery of the representative objective (one best answer).

Each of the criteria was evaluated on the following scale: to a superior degree (4), to a high degree (3), to an average degree (2), to a low degree (1), and not at all (0). Reviewers were also given space to make additional comments related to each of the five criteria and any general comments.

Reviewers worked as a group to review the three forms of each test. They were instructed to first actually take the tests (circling the correct response in the booklet) and provide comments and feedback next to each item. After reviewing all three forms in the set, each reviewer independently completed the survey asking for his or her opinion as to how well the tests met the five criteria listed above. During the last part of the session the group discussed the tests and made comments as a group. The ratings of the tests were completed anonymously. The ratings and the comments were aggregated for review by NCDPI curriculum specialists and testing consultants. The ratings for the tests are shown in Table 13.

Table 13. Average test review ratings for the North Carolina End-of-Course Tests.

Subject Area	Curricular Validity (Q #1)	Instructional Validity (Q #2)	Item Quality (Q #3)	Test/Item Bias (Q #4)	One Best Answer (Q #5)
Algebra I	3.3	2.0	2.6	3.0	2.5
Biology	3.0	2.4	1.9	2.3	2.3
ELP	2.8	3.0	2.7	2.8	2.8
English I	2.9	2.5	2.5	2.6	2.3
US History	2.8	2.0	2.1	1.9	2.2

Scales

Scores on the previous versions of the end-of-course tests were reported in terms of raw scores (number correct). The decision was made to report the new verions of the North Carolina Tests of Algebra I, Biology, ELP, English I, and US History in terms of scale scores. The end-of-course tests are not developmental in nature, so a different scale from the one used with the end-of-grade tests was needed for reporting the results. A scale with a mean of 50 and a standard deviation of 10 (range approximately 20-90) was selected. With a standard deviation of 10, one point on the scale corresponds to 0.1 of a standard deviation—typical of most standardized achievement tests.

There are many reasons to report the results from standardized tests as scale scores. Scale scores overcome the disadvantage of many other types of scores, such as percentiles, in that equal differences between score points represent equal differences in achievement. Other advanteges of scale scores are listed below.

- Scale scores can be used to compare test results when there have been changes in the curriulum and/or changes in the method of testing.
- Scale scores on pretests or released test forms can be related to scale scores used on secure test forms administered at the end of the course.
- Scale scores can be used to compare the results of tests that measure the same content area, but are composed of items presented in different formats.
- Scale scores can be used to minimize differences between various forms of the tests.

The scales for each of the end-of-course tests are calibrated to have a mean of 50 and a standard deviation of 10. Because of the observed differences in scores between field testing and statewide testing (e.g., motivation) from the end-of-grade testing program, the operational mean of the Algebra I scale was adjusted to 52. The operational means of the Biology, ELP, English I, and US History scales were adjusted to 56 because of the observed differences in scores on the Algebra I scale when it was first used statewide (e.g., the "high stakes" uses of the tests for student and school-level accountability).

Scores

Scale Scores Each student's score is determined by calculating the number of items he or she answered correctly and then converting the sum to a scale score. Because the sum of the number of items answered correctly is easy to understand and interpret, even though more sophisticated scoring options based on the IRT item parameters could be used (for example, pattern scoring), it was felt that this was the best way for the results of the North Carolina End-of-Course Tests to be reported.

The program EOC_SCAL.LSP (developed by the L.L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill) is used to convert summed scores (total number of items answered correctly) to scale scores using the three IRT parameters for each item. The scale scores produced by this program give essentially the same results as the scale scores produced by pattern scoring (easier items answered correctly count less than harder items answered correctly). Because different items are used on each form of the test, unique score conversion tables are produced for each form of the test for each subject area. For example, three forms of the Algebra I test are administered each year and, therefore, three scale score conversion tables are used in the scanning and reporting program. In addition to producing scaled scores, the program also computes the standard error of measurement associated with each summed raw score. See Appendix E for further information concerning the methodology used to convert summed scores to scaled scores.

Percentile Ranks In addition to scale scores, the percentile ranks associated with each scale score within each subject area are also reported at the individual level. The percentile rank for each scale score is the percentage of students in the course who obtained scores lower than that scale score. The percentile ranks provide relative information on the performance of students.

The percentile ranks for the scores on the North Carolina End-of-Course Test of Algebra I were calculated based on the May 1994 administration of the test. The percentile ranks for the scores on the North Carolina End-of-Course Tests of Biology, English I, and US History were calculated based on the May 1995 administrations of the tests. Estimated percentile ranks for scores on the North Carolina End-of-Course Test of ELP, derived from the EOC_SCAL.LSP program, will be reported during the 1996-97 school year. Final percentile ranks for the ELP test will be calculated based on all administrations of the test during the 1996-97 school year. The percentile tables are published in *State Norms Tables for North Carolina* (NCDPI, 1995).

Descriptive Statistics

The North Carolina End-of-Course Test of Algebra I was first administered in May 1994 and then throughout the 1994-1995 school year—Forms A, B, and C. The test booklets were spiraled within each classroom, such that one-third of the students received form A, one-third of the students received form B, and one-third of the students received form C. Three additional equivalent forms—Forms G, H, and I—were administered for the first time in the Summer and Fall of 1995. The North Carolina End-of-Course Tests of Biology, English I, and US History were first administered in May 1994—Forms A, B, and C. Three additional equivalent forms of each of the tests—Forms D, E, and F—were administered for the first time in the Summer and Fall of 1995. The North Carolina End-of-Course Test of Economic, Legal, and Political Systems (ELP) will be administered for the first time in the Summer of 1996—Forms A, B, and C. Table 14 presents the descriptive statistics from the first administration of the tests in May 1994 and the second administration of the Algebra I test during the 1994-1995 school year.

Table 14. Descriptive Statistics for the first administration of the North Carolina End-of-Course Tests—Forms A, B, and C.

Subject Area	N	Mean Scale Score	Mean by Form (Range)	Standard Deviation
Algebra I				
1994	69,326	55.1	54.1–55.5	9.1
1995	80,370	54.7	54.6–55.7	9.4
Biology	62,480	55.5	54.7–56.1	8.6
English I	67,748	53.1	52.6–53.6	8.9
US History	53,160	56.2	56.1–56.4	8.3

Of special significance to the comparison of student scores across time, and scores in general across time, is the equivalence of the test forms. All forms developed for a subject area (for example, English I, Forms A, B, and C) were equated to the mean derived from the latent proficiency scaling of the tests using the EOG_SCAL.LSP program. From an examination of Table 14 above, the differences between the mean scores across the forms are small and are always less than the standard error of measurement for the test (see Table 16).

Figures 6 through 9 present the frequency distributions of the developmental scale scores from the first administration of the tests. The frequency distributions are not smooth because of the conversion from raw scores to scale scores. Due to rounding in the conversion process, sometimes two raw scores in the middle of the distribution convert to the same scale score.

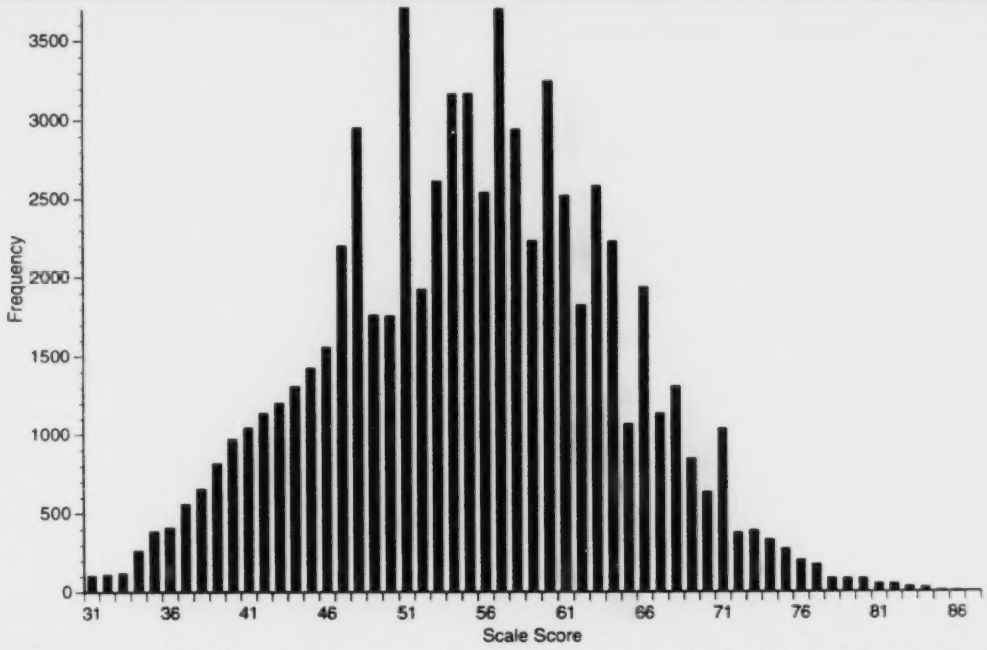


Figure 6. Frequency distribution of scores on the North Carolina End-of-Course Test of Algebra I—Forms A, B, and C (N = 69,162).

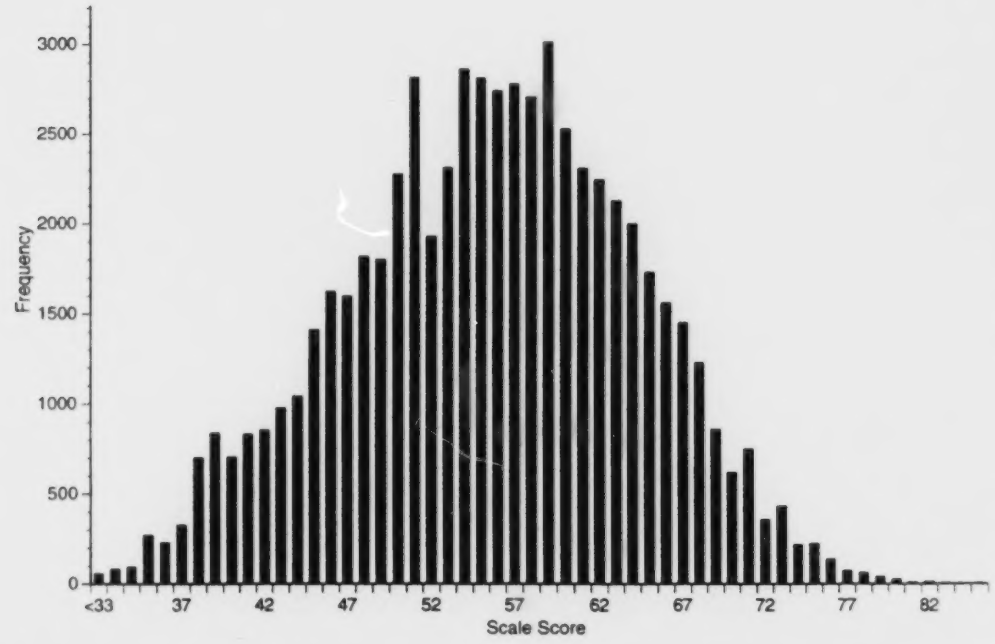


Figure 7. Frequency distribution of scores on the North Carolina End-of-Course Test of Biology—Forms A, B, and C (N = 62,470).

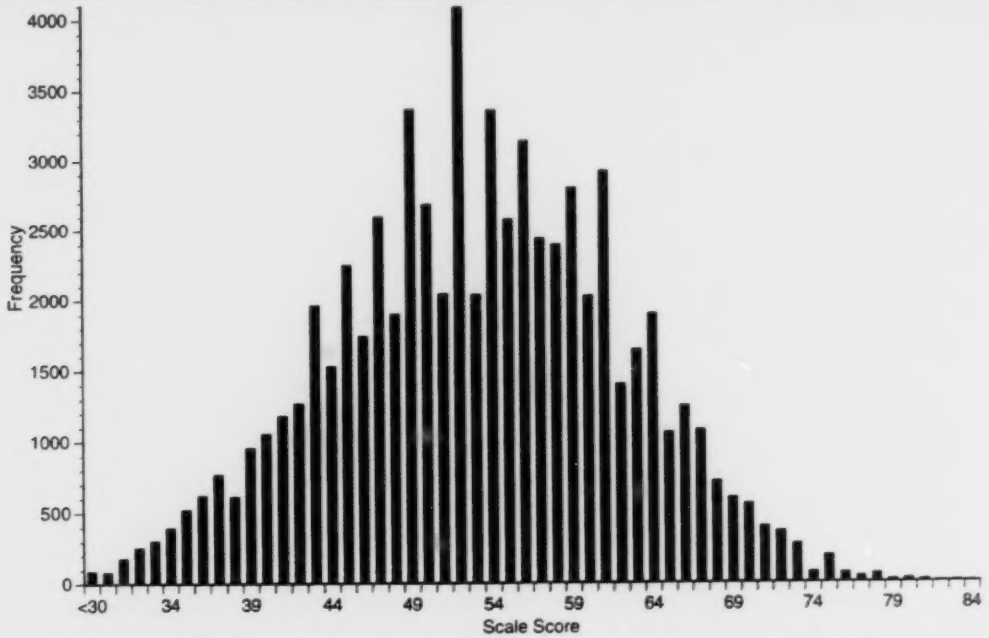


Figure 8. Frequency distribution of scores on the North Carolina End-of-Course Test of English I—Forms A, B, and C (N = 67,739).

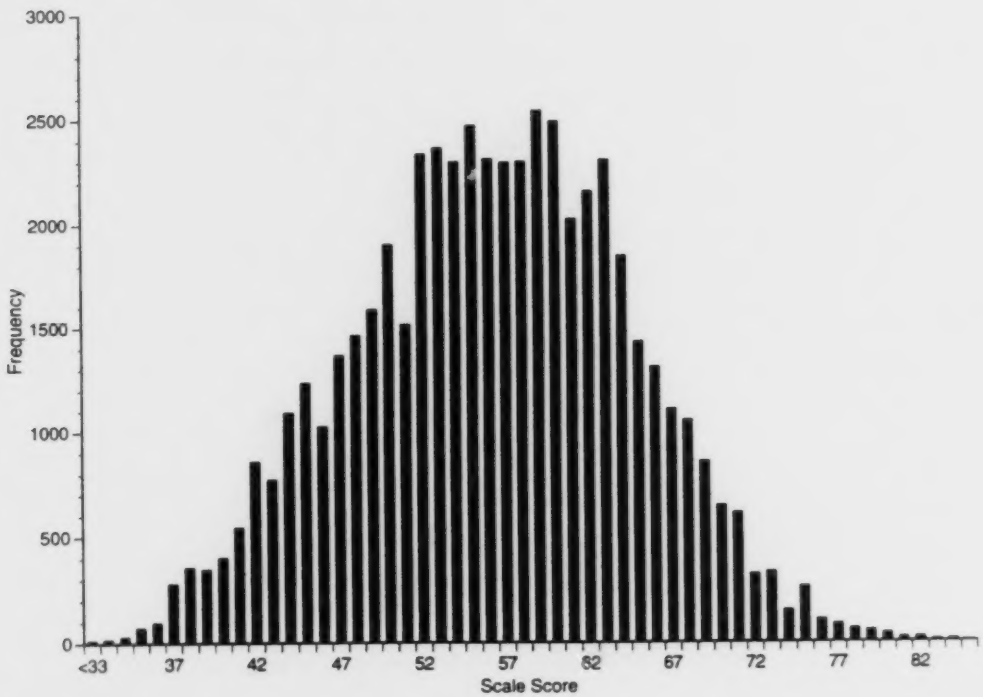


Figure 9. Frequency distribution of scores on the North Carolina End-of-Course Test of US History—Forms A, B, and C (N = 53,157).

Reliability

Reliability refers to the consistency of scores obtained by the same person when examined with the same test on different occasions or with different sets of equivalent items. If any use is to be made of the information from a test, then it is desirable that the test results be reliable. If decisions about individuals are to be made on the basis of the test data (for example, placement or instructional program decisions), then it is desirable that the test results be reliable and exhibit a reliability coefficient of at least 0.85. In testing, if use is to be made of some piece of information, then the information should be stable, consistent, and dependable.

Internal-Consistency Reliability Internal-consistency reliability examines the extent to which the test measures a single basic concept. One procedure for determining the internal consistency of a test is coefficient alpha (α). Coefficient alpha sets an upper limit to the reliability of tests constructed in terms of the domain-sampling model. Table 15 presents the item-level values of coefficient α for the North Carolina Tests of Algebra I, Biology, English I, and US History. Coefficient α estimates for the pretests were determined by the Spearman-Brown Prophecy Formula for a test half as long as the original test.

Table 15. Item-level values of coefficient α for the first administration of the North Carolina End-of-Course Tests (Forms A, B, and C) and pretests.

Subject Area	Regular EOC Test	EOC Pretest
Algebra I	0.94	0.89
Biology	0.94	0.88
English I	0.92	0.84
US History	0.92	0.85

Standard Error of Measurement The standard error of measurement of a test is the standard deviation of

$$\sigma_{meas} = SD_t\sqrt{1-r_{tt}} \qquad \text{(Equation 4)}$$

the error scores of a test. Typically the standard error of measurement is determined by the following formula: When using item response theory measurement, the test information function, which depends only on the items included in the test, permits the estimation of the error of measurement at each ability level (or score).

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \qquad \text{(Equation 5)}$$

The standard error of measurement (estimation in IRT methodology) is determined by the following formula: The magnitude of the standard error of θ (an examinee's estimated ability level) depends on the following characteristics of the test:

- the number of test items—smaller standard errors are associated with longer tests,
- the quality of the test items—in general, smaller standard errors are associated with highly discriminating items for which the correct answers cannot be obtained by guessing, and
- the match between item difficulty and examinee ability—smaller standard errors are associated with tests composed of items with difficulty parameters approximately equal to the the ability parameter of the examinee, as opposed to tests that are relatively easy or relatively difficult (Hambleton, Swaminathan, and Rogers, 1991).

Table 16 shows the standard error of measurement ranges for scores on the North Carolina End-of-Course Tests. For students with scores within two standard deviations of the mean (95% of the students), standard errors are typically 2 to 3 points. As scores become more extreme there is less measurement precision associated with a score.

Table 16. Standard error of measurement for ranges of scores on the North Carolina End-of-Course Tests.

Score Range	Algebra I	Biology	ELP	English I	US History
80–89	4	3	4	4	3
70–79	3	2	3	3	3
60–69	2	2	2	2	2
50–59	3	2	2	3	2
40–49	4	4	3	4	4
30–39	6	6	5	5	6
20–29	5	5	5	5	5

The validity of a test is the degree to which the test actually measures what it purports to measure. Validity provides a direct check on how well the test fulfills its function. For all forms of test development, validity is a predominant theme from the time the idea for the test is conceived until the final test scores have been analyzed and interpreted. For convenience, the various components of test validity—content, criterion-related, and construct—will be described as if they were unique, independent components rather than interrelated parts.

Content Validity

The content validity of a test relates to the adequacy with which important content has been sampled and the adequacy with which the content is evidenced in the test items. Content validity was built into the North Carolina End-of-Course Tests during the development process. All items are aligned with the North Carolina *Standard Course of Study* for Mathematics, English Language Arts, Science, and Social Studies, the basis for instruction in North Carolina schools. The items were written and reviewed by North Carolina teachers who are in contact with the students every day in the classroom.

Criterion-Related Validity

Criterion-related validity of a test indicates the effectiveness of a test in predicting an individual's behavior in a specific situation. The criterion for evaluating the performance of the test can be measured at the same time (concurrent validity) or at some later time (predictive validity). The following discussion of the relationship between scores on the North Carolina End-of-Course Tests and teacher judgments of student achievement is evidence of concurrent validity.

Achievement Levels The North Carolina *Standard Course of Study* outlines the content standards for North Carolina in that it describes the knowledge, skills, and other understanding that schools should teach in order for students to attain high levels of competency in challenging subject matter. Educators in North Carolina felt that performance standards should also be developed which identify levels of competency expected in each content area. Unlike percentiles, which yield only relative comparisons, the performance standards give common meaning throughout the state as to what is expected at various levels of competence in each subject area. Performance standards, called Achievement Levels, are one way that scores on the North Carolina End-of-Course Tests are reported. These categories are used to better describe the scores on the tests and are based on external evidence about the relative skill of students.

The achievement levels for the North Carolina End-of-Course Tests are based on the contrasting groups method of standard setting. This method involves having students categorized into the various achievement levels by expert judges who are knowledgeable of the students' achievement in various domains assessed outside of the testing situation. Teachers are able to make informed judgments about students' achievement because the teachers have observed the breadth and depth of the work each student has accomplished during the school year.

During field testing, teachers were asked to categorize each of their students on the basis of "absolute" achievement (comparison to an external standard). Each student was categorized into one of four achievement levels based on the teacher's experiences with the student throughout the school year.

Level I	Students performing at this level *do not have sufficient mastery* of knowledge and skills in [name of subject area] to be successful at a more advanced level in the content area.
Level II	Students performing at this level *demonstrate inconsistent mastery* of knowledge and skills in [name of subject area] and are minimally prepared to be successful at a more advanced level in the content area.
Level III	Students performing at this level *consistently demonstrate mastery* of [name of subject area] subject matter and skills and are well prepared for a more advanced level in the content area.
Level IV	Students performing at this level *consistently perform in a superior manner* clearly beyond that required to be proficient in [name of subject area] subject matter and skills and are very well prepared for a more advanced level in the content area.
	or
	Not a clear example of any of these achievement levels.

In all, the judgments of more than 650 teachers about the performance of approximately 50,000 students were involved in the standard setting process statewide. About 80% of the students field tested were categorized into one of the four achievement levels, with the remainder categorized as not a clear example of any of the achievement levels or not categorized at all.

The percentage of students in each achievement level were remarkably similar across subjects. The percentages are presented in Table 17 below.

Table 17. Percent of students assigned to each achievement level by teachers during field testing.

Subject Area	Level I	Level II	Level III	Level IV
Algebra I	14.5%	32.5%	40.4%	12.6%
Biology	17.3%	30.9%	36.3%	15.4%
ELP	13.7%	27.1%	36.0%	23.2%
English I	13.4%	32.3%	35.4%	18.9%
US History	17.3%	33.6%	33.6%	15.4%

Figure 10 shows the relationship between students' scores on the field test with the teacher judgements concerning achievement. As expected, the percent of correct items on the test increases over the achievement levels. Students rated by teachers as high achievers (Level IV) scored high on the tests, while students who were rated low by teachers scored low on the test (Level I).

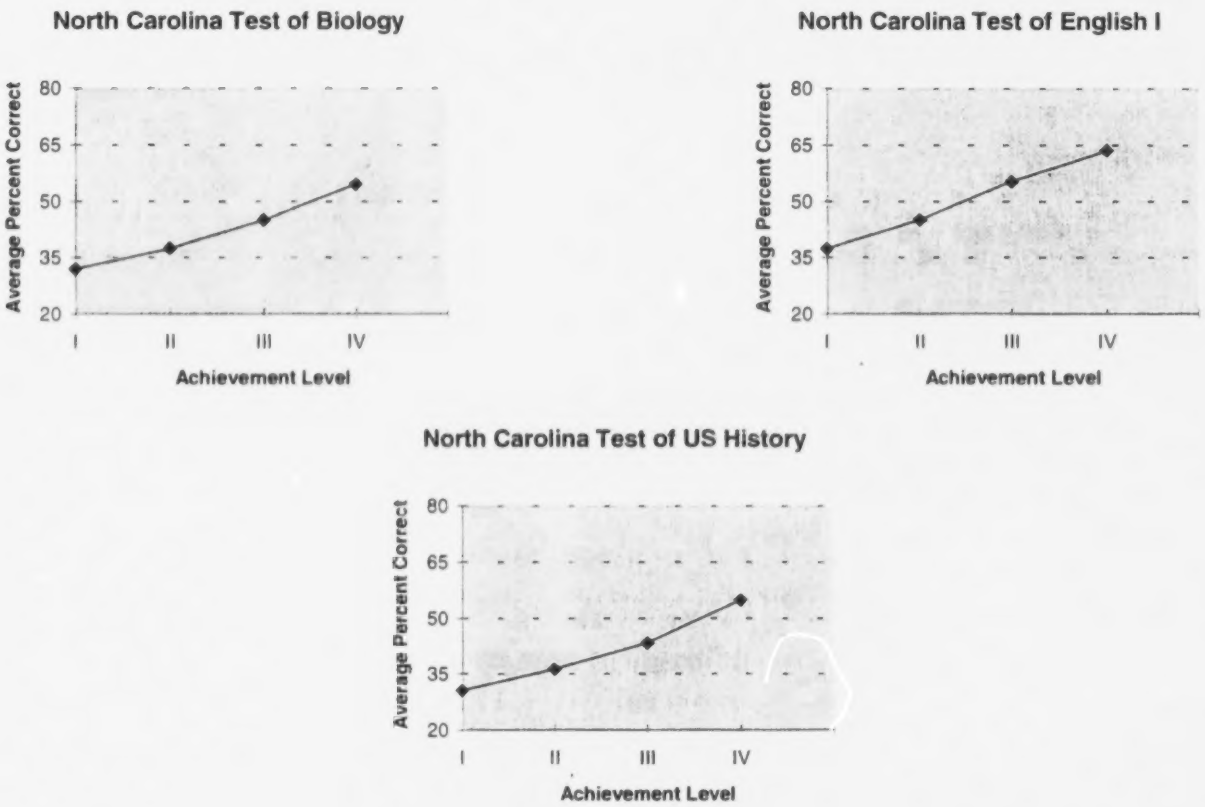


Figure 10. The relationship between teacher judgements of student achievement and scores on the North Carolina End-of-Course Tests during field tesing.

The percentages of students shown in Table 17 for each subject area were used in conjunction with the frequency distributions of scores from the first administration of the North Carolina End-of-Course Tests to determine where the cut points should be for the achievement levels. Table 18 gives the range of scores associated with each achievement level. The range of scores associated with the ELP achievement levels are based on the estimated distribution of scores and will be revised once the test has actually been administered and the data has been aggregated for state-level analyses.

Table 18. Range of scores associated with each achievement level for score reporting.

Subject	Level I	Level II	Level III	Level IV
Algebra I	23–44	45–54	55–65	66–87
Biology	26–46	47–54	55–64	65–87
ELP	23–44	45–53	53–62	63–85
English I	22–42	43–51	52–60	61–85
US History	27–47	48–56	57–64	65–88

Note: ELP score ranges based on estimated distribution of scores.

Construct Validity

The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait, such as reading comprehension or mathematics achievement. "Correlations between a new test and other similar tests . . . [are] evidence that the new test measures approximately the same general area of behavior as other tests designated by the same name" (Anastasi, 1982). The following sections provide evidence of the construct validity of the North Carolina End-of-Course Tests of Algebra I, Biology, English I, and US History.

North Carolina Open-Ended Tests The North Carolina Open-Ended Tests of Biology, English I, and US History were designed to measure broadly higher level thinking skills by requiring students to apply or demonstrate skills and knowledge beyond the recall level. These items were designed to be open so that the quality of the student's response would determine his or her score. Some of the items required the production of a specific answer to a problem, but the student was also asked to explain how he or she arrived at the answer. This explanation helped to determine the student's score.

The open-ended items were written and reviewed by advisory committees composed of testing consultants, teachers, curriculum specialists, and university professors. The items were field tested on approximately 1,000 students randomly selected from across the state to examine how each item performed (score distribution) and to refine the scoring rubric. Specific scoring rubrics, based on the general scoring rubric for the content area, describe the standards used to judge specific items within the content area. The items were field tested a second time to verify the scoring rubric. Results were analyzed using Samejima's graded item response theory model. The field test information was examined for each item and the decision was made to retain the item for future test development or to delete the item at this time (Hambleton, Swaminathan, & Rogers, 1991).

The Biology, English I, and US History items were field tested in March 1994. The same students were administered an open-ended field test and a multiple-choice field test in May. For all three subject areas, each open-ended field test form consisted of four items and reflected the breadth and depth of the curriculum.

For Biology, the correlations between multiple-choice total scores and scores on the open-ended items ranged from 0.24 to 0.51 (mean = 0.40) with approximately 200 students per matched field test form. For English I, only the first three open-ended items were used in this analysis; the fourth item asked the student to reflect on the selections he or she had read during the year and relate one of the selections to the theme of the field test form or to some literary characteristic of the passage in the field test form. The correlations between the multiple-choice total scores and scores on the open-ended items ranged from 0.32 to 0.57 (mean = 0.44) with approximately 175 students per matched field test form. For US History, the correlations between multiple-choice total scores and scores on the open-ended items ranged from 0.28 to 0.55 (mean = 0.43) with approximately 150 students per matched field test form. Considering the differences in format, this is the level of correlation that would be expected between the two sets of scores for each subject area in a multi-method design (Campbell and Fiske, 1959).

North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics The North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics are achievement tests developed by the NCDPI for the assessment of achievement at the end of each grade, 3 through 8. The reading comprehension test assesses the same goals and objectives as the end-of-course English I test assesses; the only difference between the two tests is that the English I test has a greater emphasis on literary analysis and a lesser emphasis on reading comprehension. The English I test also assesses the student's ability to edit for grammar, mechanics, usage, and spelling. The Algebra I test assesses goals 3 (pre-algebra), 5 (problem solving), 6 (statistics and data analysis), and 7 (computation) of the mathematics *Standard Course of Study* for grades 3 through 8. For information concerning the psychometric characteristics of the reading comprehension and mathematics tests for grades 3 through 8 see *Technical Characteristics of the North Carolina End-of-Grade Tests* (NCDPI, 1996).

The scores of students who were administered the 1995 North Carolina End-of-Course Tests of English I and Algebra I were matched with the their scores on the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics in Grade 8. A total of 43,194 students had both English I and end-of-grade reading test scores. The correlation between the two sets of scores was 0.81. A total of 27,076 students had both Algebra I and end-of-grade mathematics test scores. The correlation between the two sets of scores was 0.73.

Expected Grades When the Algebra I, Biology, English I, and US History were first administered statewide, additional student information was collected. Teachers were asked to indicate the excpected final letter grade for each student in their classess. Figure 11 dispalys a comparison of letter grades in each course and the mean end-of-course test score corresponding to each letter grade for the overall student population. The Algebra I results are from the Spring 1995 administration of the test to 68,581 students; the correlation between expected grade and test score was 0.62. The Biology results are from the Spring 1995 administration of the test to 62,241 students; the correlation between expected grade and test score was 0.54. The English I results are from the Spring 1995 administration of the test to 67,213 students; the correlation between expected grade and test score was 0.54. The US History results are from the Spring 1995 administration of the test to 52,929 students; the correlation between expected grade and test score was 0.53.

Both the graphs and the correlations correspond closely to expectation and provide evidence concerning the validity of the tests.

Laboratory Skills On the field test of the North Carolina Test of Biology, students were asked how often they do labs in Biology class. A total of 8,780 students responded to the question. Based on the results in Table 19,

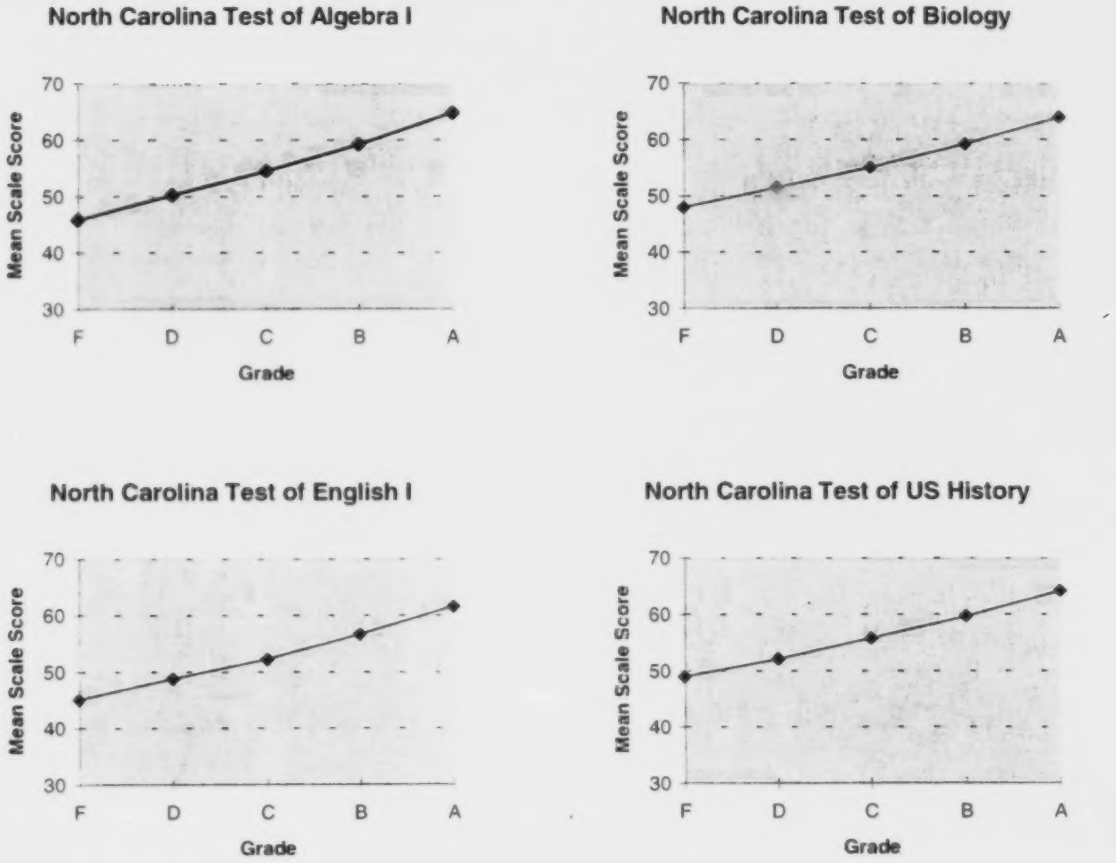


Figure 11. Comparison of letter grades teachers expected students to receive and scores subsequently earned on the North Carolina End-of-Course Tests.

students typically do laboratory experiences in Biology two to three times a month. In addition, the results in the table show that as students participate in more laboratory experiences, their scores on the Biology test increase.

Table 19. Results from Survey Question on Laboratory Experience and the North Carolina Field Test of Biology.

How often do you do labs?	Percent of Students Responding	Mean Percent Correct
Once a Week	12.8%	50%
2-3 Times per Month	42.2%	44%
Once a Month	23.5%	42%
Seldom	18.2%	39%
Never	3.2%	36%

Anastasi, A. (1982). *Psychological Testing*. New York: Macmillan Publishing Company, Inc.

Averett, C.P. (1994). North Carolina End-of-Grade Tests: Setting standards for the achievement levels. Unpublished manuscript.

Berk, R.A. (1984). *A Guide to Criterion-Referenced Test Construction*. Baltimore: The Johns Hopkins University Press.

Berk, R.A. (1982). *Handbook of Methods for Detecting Test Bias*. Baltimore: The Johns Hopkins University Press.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full information factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage Publications, Inc.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cattell, R.B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology, 12*, 105-214.

Dorans, N.J. & Holland, P.W. (1993). DIF Detection and description: Mantel-Haenszel and standardization. In P.W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (pp 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Haladyna, T.M. (1994). *Developing and Validating Multiple-Choice Test Items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.

Holland, P.W. & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Joreskog, K.J. & Sorbom, D. (1986). *PRELIS: A program for multivariate data screening and data summarization*. Chicago, IL: Scientific Software, Inc.

Joreskog, K.J. & Sorbom, D. (1988). *LISREL 7: A guide to the program and applications*. Chicago, IL: SPSS, Inc.

Kubiszyn, T. & Borich, G. (1990). *Educational Testing and Measurement*. New York: HarperCollins Publishers.

Marzano, R.J., Brandt, R.S., Hughes, C.S., Jones, B.F., Presseisen, B.Z., Stuart, C./, & Suhor, C. (1988). *Dimensions of Thinking*. Alexandria, VA: Association for Supervision and Curriculum Development.

BEST COPY AVAILABLE

Michigan Curriculum Review Committee and Michigan Reading Association. (1985). *New Dimensions in Reading: Instruction.*

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Muraki, E., Mislevy, R.J., & Bock, R.D. *PC-Bimain Manual.* (1991). Chicago, IL: Scientific Software, Inc.

National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics.* (1989). Reston, VA: Author.

North Carolina Department of Public Instruction. (1992). *Teacher Handbook for English Language Arts.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1992). *Teacher Handbook for Mathematics.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1992). *Social Studies Standard Course of Study.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1993). *North Carolina End-of-Grade Testing Program: Background Information.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1994). *Science Standard Course of Study.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1996). *North Carolina Testing Code of Ethics.* Raleigh, NC: Author.

North Carolina State Board of Education. (1993). *Public School Laws of North Carolina 1994.* Raleigh, NC: The Michie Company.

North Carolina State Board of Education. (March 1996). *Examining the Structure and Functions of the Public School System of North Carolina.* Raleigh, NC: Author.

Nunnally, J. (1978). *Psychometric Theory.* New York: McGraw-Hill Book Company.

Rosenthal, R. & Rosnow, R.L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill Book Company.

SAS Institute, Inc. (1985). The FREQ Procedure. In *SAS User's Guide: Statistics, Version 5 Edition.* Cary, NC: Author.

Schmid, J. & Leiman, J.M. (1957). The devlopment of heirarchical factor solutions. *Psychometrika, 22,* 53-61.

Traub, R.E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage Publciations, Inc.

Appendices

Appendix A: Sample Items and Passages with Curricular and Psychometric Information A–1

Appendix B: Reading and Mathematics Curricula, Test Specifications, and Average Difficulty of Item Pools B–1

Appendix C: North Carolina Technical Advisory Committee C–1

Appendix D: Sample Item with Development, Review, and Psychometric Information D–1

Appendix E: *Excerpt from* "Item Response Theory for Scores on Tests including Polychotomus Items with Ordered Responses" E–1

Appendix F: Bias Review Materials F–1

Appendix A

Sample Items and Passages with Curricular and Psychometric Information

Algebra I A–2

Biology A–11

Economic, Legal, and Political Systems A–18

English I (Editing and Textual Analysis) A–27

US History A–38

Released Items

Algebra I Goal 3

Objective 3.3 — Solve an equation graphically and by using more than one property of equality.

Which pair of equality properties would be used to solve $3x - 5 = 7$?

A subtraction by 7 and multiplication by 3

B multiplication by $\frac{1}{3}$ and addition by 7

C division by 5 and subtraction by 7

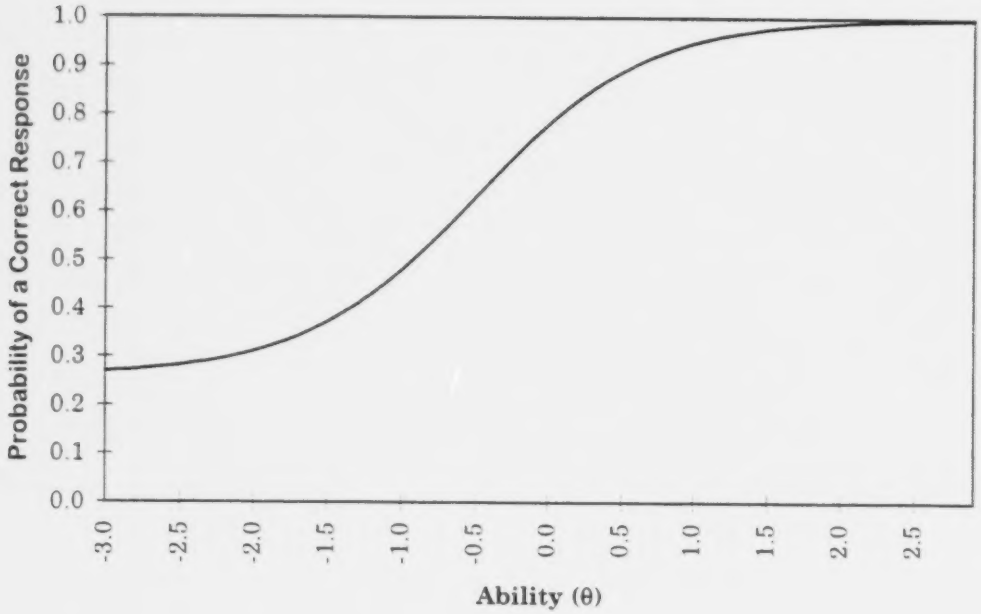
D addition by 5 and division by 3

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
1831	2	20	3.03	4	83	46	56	547

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.74	0.532	.44	1.202	1.519	-0.499	1.012	.259

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

Algebra I Goal 4

Objective 4.10 — Compare ordered pairs to the line $y = x$ and interpret the results.

Five students begin training for outdoor track. Prior to the training program, the students' weights were 135, 147, 193, 171, and 168 pounds. After the training program, the students' weights were 134, 151, 190, 173, and 165 respectively. How many of the ordered pairs in this data set fall on, above, or below the line $y = x$?

A 3 above, 2 below

B 1 on, 2 above, 2 below

C 2 above, 3 below

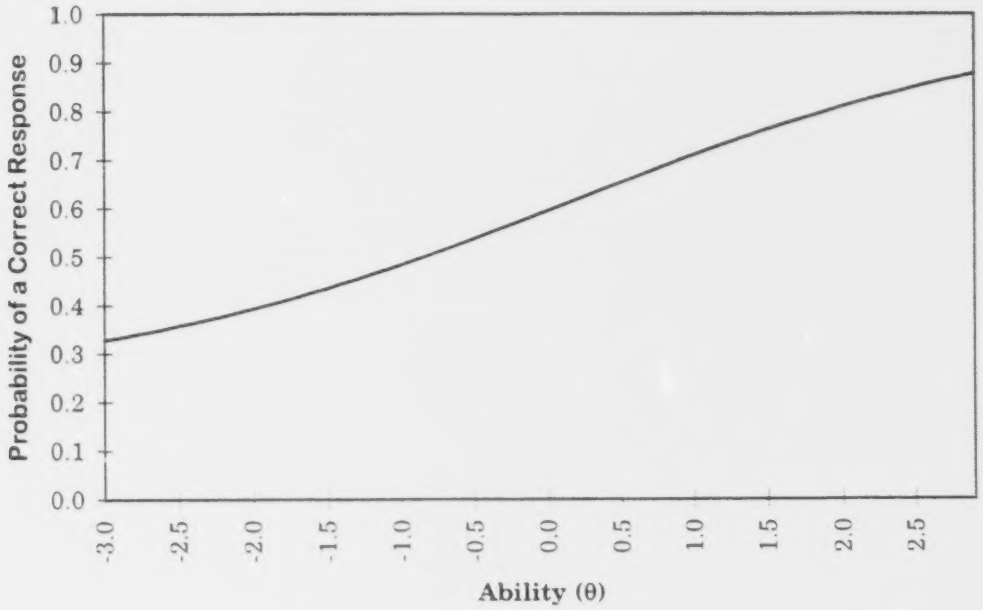
D 1 on, 1 above, 3 below

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2098	3	41	4.10	3	116	103	431	47

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.60	0.227	.49	1.187	0.849	0.168	0.357	.229

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

Algebra I Goal 8

Objective 8.3 — Use ratios and proportions to solve problems.

If a 16-meter tree casts a 7.5-meter shadow, how tall is a building that casts a 24-meter shadow at the same time of day?

A 11.2 m

B 50 m

C 51.2 m

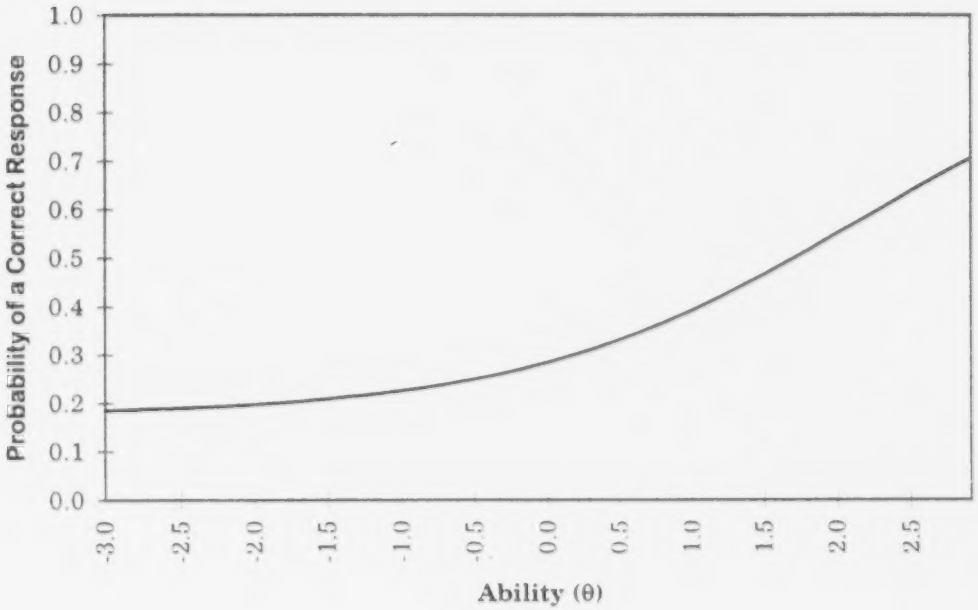
D 512 m

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2338	5	77	8.3	3	189	146	220	58

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.31	0.239	.46	0.841	0.797	2.217	0.502	.175

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

Algebra I Goal 9

Objective 9.10 — Determine if a set of data represents an exponential function.

The frequencies (cycles per second) of the sound of A's on a piano increase in an exponential fashion as you move up the keyboard. Which of the following are frequencies of key A?

A 55, 110, 440, 3520, 56320

B 55, 110, 220, 440, 880

C 55, 110, 165, 220, 275

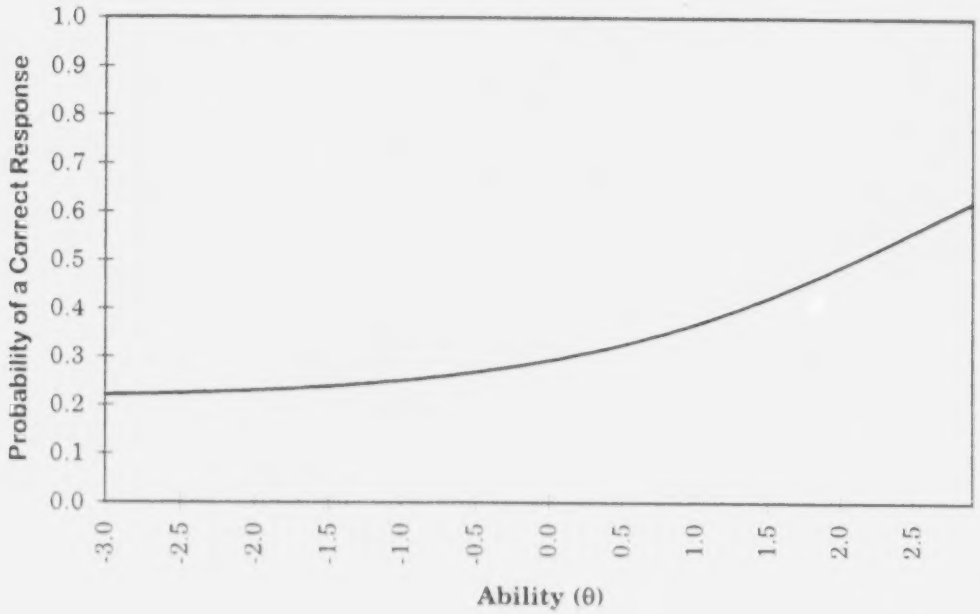
D 55, 105, 155, 205, 255

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2474	4	87	9.10	2	85	224	180	104

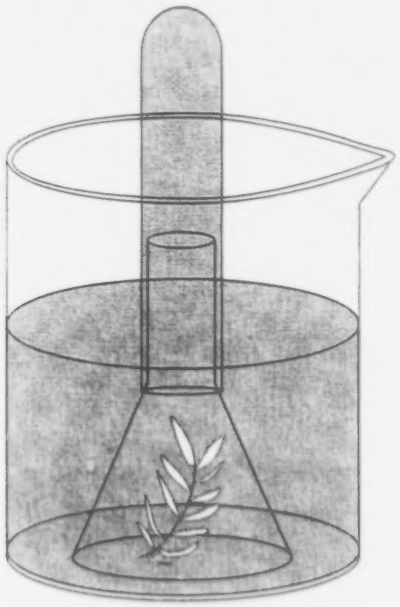
			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.31	0.160	.46	0.723	0.947	2.796	0.455	.211

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Student Directions: Use the following information to answer the next two questions.

In the drawing below, an *Elodea* plant is placed in the bottom of a beaker that is half filled with water. A funnel is placed over the plant. A test tube filled with water is inverted and placed on top of the funnel. The beaker is placed in direct sunlight. The purpose of the experiment is to determine the rate of photosynthesis in the *Elodea* plant.



BEST COPY AVAILABLE

Released Items

Biology Goal 6

Objective 6.3 — Demonstrate knowledge of the chemical processes of life.

How could this experiment show that photosynthesis is taking place?

A The *Elodea* will turn green.

B Chlorophyll will dissolve in the water.

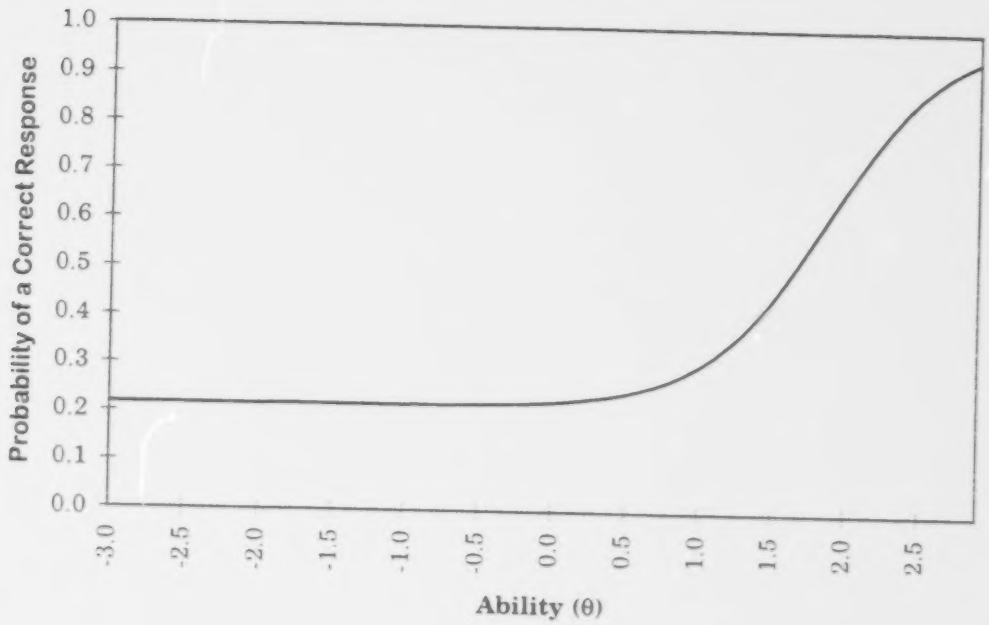
C Bubbles will appear in the test tube.

D New leaves will grow on the *Elodea*.

Item Statistics

						--------Choice--------			
Origno	Form	Item	Obj	SG	Key	A	B	C	D
4105	2	13	6.3	03	3	216	197	251	258
			-----Bias-----			------IRT Parameters------			
P	Bis	Psd	Ethnic	Gender		Threshold	Slope	Asymptote	
.27	0.27	.44	1.016	1.516		1.86577	1.4289	.21745	

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

Biology Goal 6

Objective 6.3 — Demonstrate knowledge of the chemical processes of life.

If you were to set up a control for this experiment, what would be the ***best*** setup?

A the same setup but no water

B the same setup but no plant

C the same setup but a larger plant

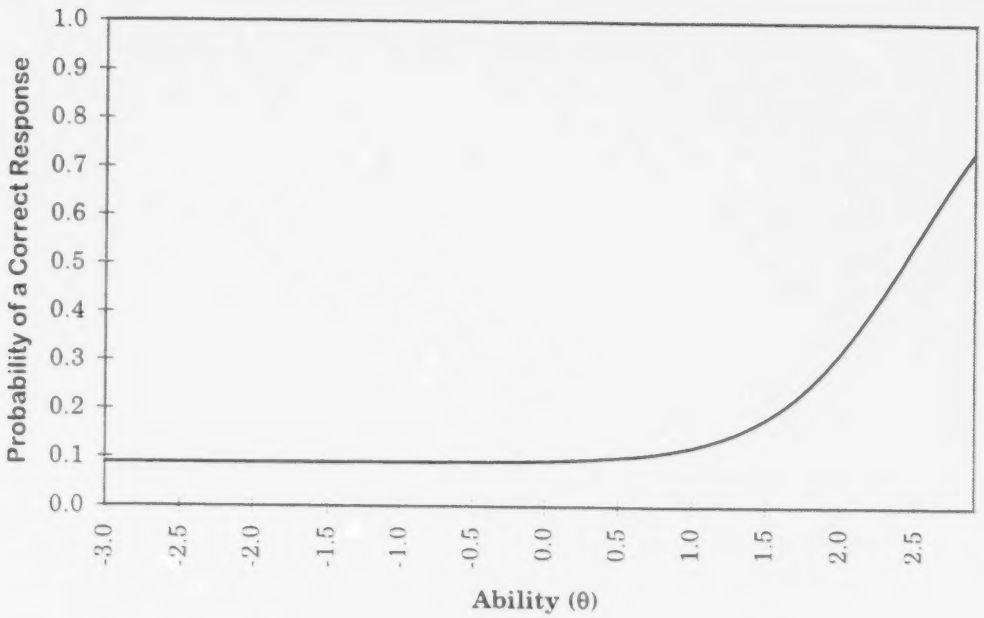
D the same setup but no plant and no direct sunlight

Item Statistics

						--------Choice--------			
Origno	Form	Item	Obj	SG	Key	A	B	C	D
4106	2	14	6.3	02	2	418	97	285	120

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.1	0.108	.31	.947	1.739	2.49989	1.29779	.08996

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

Biology Goal 6

Objective 6.3 — Demonstrate knowledge of the chemical processes of life.

Which of the following relationships is ***most*** similar to perspiration : pores?

A epidermis : stomata

B circulation : parenchyma

C digestion : chlorophyll

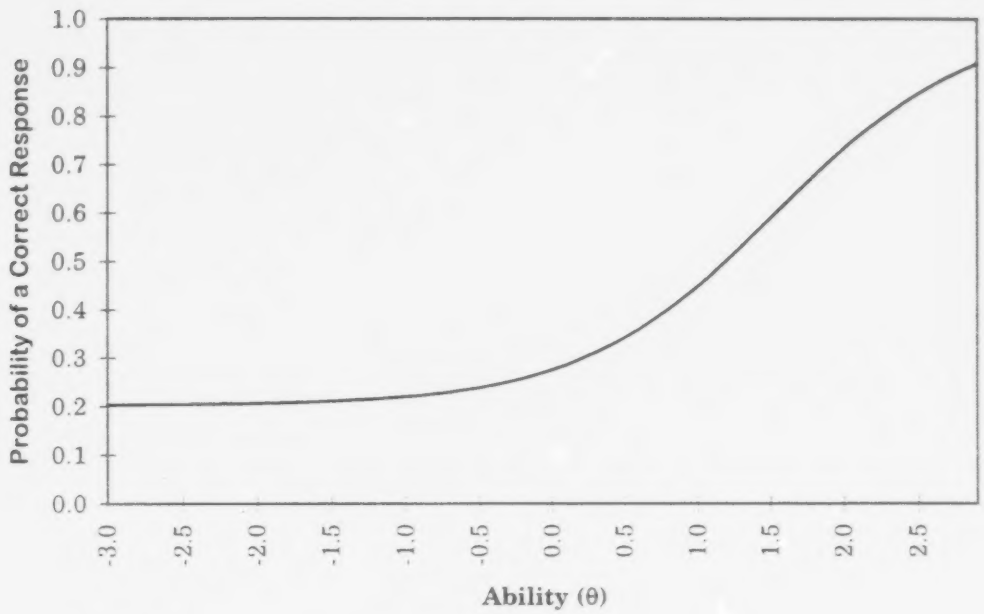
D transpiration : stomata

Item Statistics

						--------Choice--------			
Origno	Form	Item	Obj	SG	Key	A	B	C	D
4174	2	31	6.3	02	4	248	197	165	305

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.33	0.363	.47	1.005	0.872	1.54072	0.8778	.20243

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

ELP Goal 1

Objective 1.1 — Know examples of recurring public problems and issues.

For which reason would a legislator from North Carolina be ***most likely*** to argue against a proposed tax on tobacco?

A North Carolinians are traditionally opposed to "sin" taxes.

B The tax could mean smaller returns on investment for tobacco producers in North Carolina.

C The tax could mean that fewer low-income people in North Carolina could afford tobacco products.

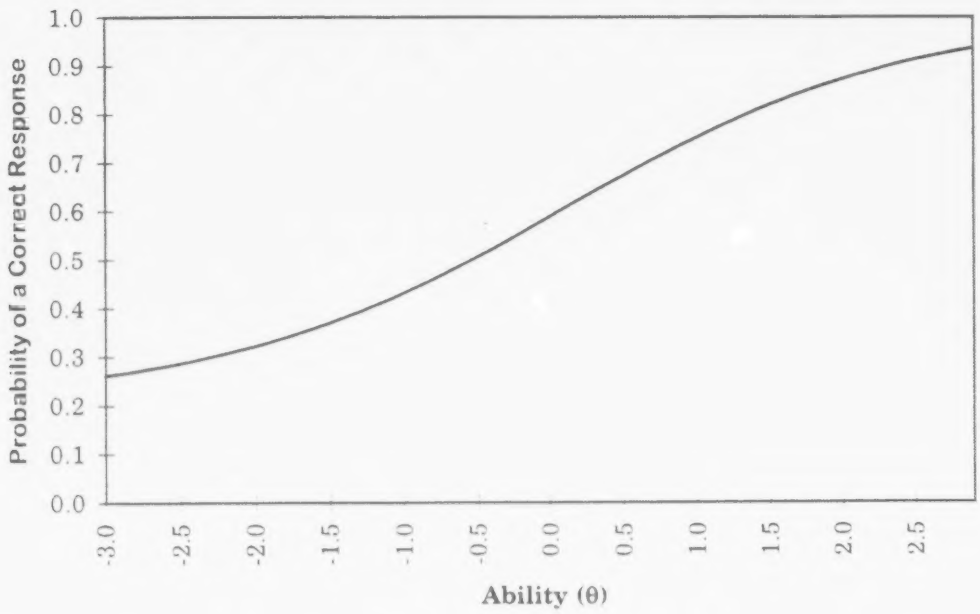
D North Carolinians are proud of the contributions made by tobacco throughout the state's history.

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
R001	11	2	1.1	2	39	467	172	110

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.59	0.49	.351	1.323	0.924	0.1	0.508	.21

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

ELP Goal 4

Objective 4.4 — Demonstrate how supply, demand and competition affect prices and the availability of goods and services.

Which of the following events will increase the demand for compact cars?

A An oil shortage pushes up the cost of gasoline.

B Slow sales cause General Motors to cut luxury car prices.

C More thefts lead to higher automobile insurance premiums.

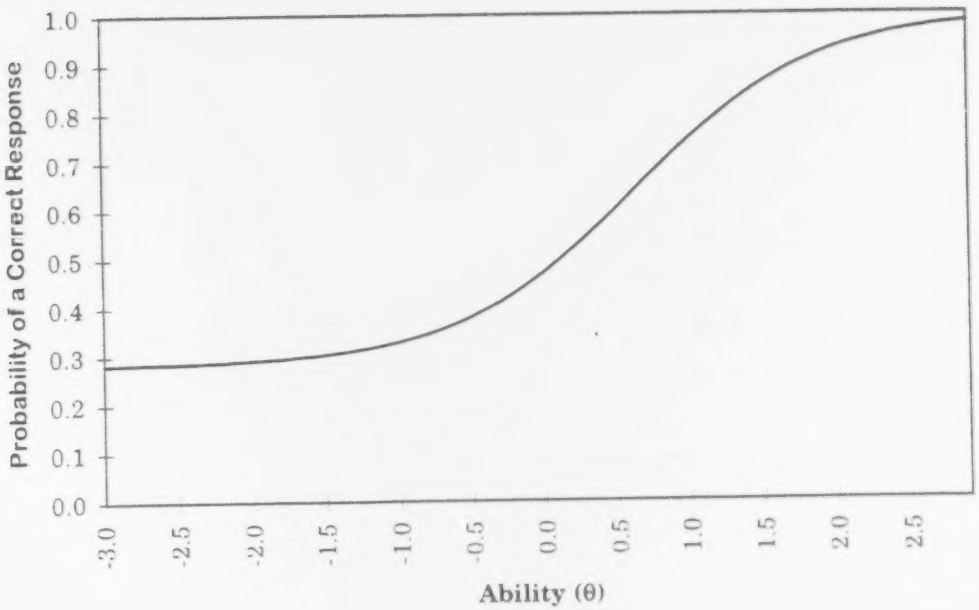
D Concern for the environment leads to an increase in carpooling.

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
R004	11	29	4.4	1	412	156	104	112

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.52	0.446	.5	1.611	1.36	0.64	0.952	.28

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

ELP Goal 5

Objective 5.5 — Analyze short and long term effects of taxation and government spending on the United States economy.

In 1790, Alexander Hamilton, the first Secretary of the Treasury, felt that the finances of the new Republic were in a state of chaos. He thought that the best way to introduce "order into our finances" would be to have the powerful and wealthy invest in the country, while letting the common people fend for themselves.

Based on Hamilton's views in the 1790s which current trend would ***most*** disturb him?

A the government's refusal to benefit the farmers by placing high tariffs on agricultural products

B decreased lending by the government to entrepreneurs for small businesses

C increased spending by the government to provide services for the poor

D the government's inability to develop a fair and equitable tax plan for the benefit of all citizens

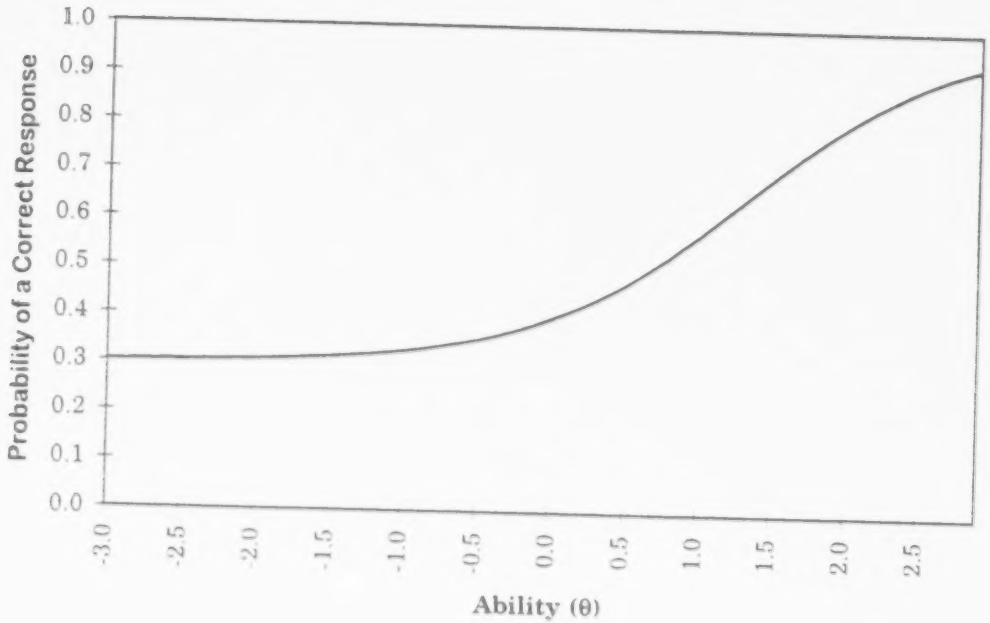
---

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
R008	8	42	5.5	3	135	137	360	157

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.45	0.322	0.5	0.996	0.872	1.32	0.813	0.3

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

ELP Goal 8

Objective 8.6 — Understand the responsibility, jurisdiction and methods of individual law-enforcement agencies.

A preliminary hearing is usually held after which legal action?

A sentencing

B arrest

C verdict by jury

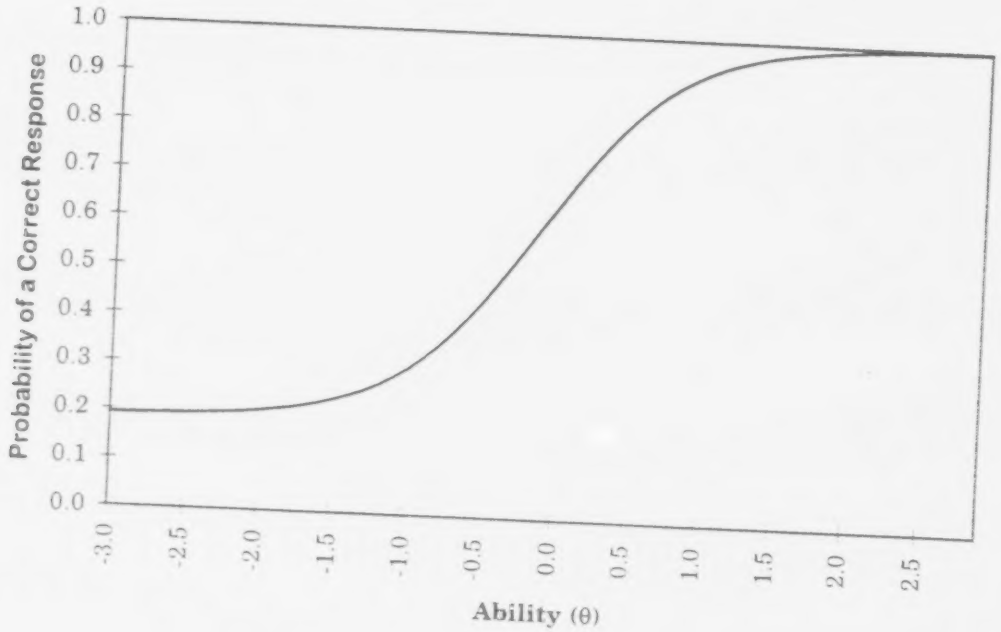
D parole

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
R005	5	73	8.6	2	100	506	79	80

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.63	0.651	0.48	0.844	0.847	-0.13	1.243	0.19

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Student Directions: Use the following student essay to answer the next two questions.

In the novel *Night*, by Elie Wiesel, the author shows how the atmosphere of the outside world affects the inner feelings and attitudes of those who endure suffering. Thorough (1) his words, the author paints a graphic picture of how his (2) living during the massacre of the Jews affected his faith and his will to survive in a time of endless death.

The author lives in a hateful world of merciless killings; the (3) horror tests his faith, not only in himself but also in the God he has worshipped. For long periods during his life, he being taught (4) that his God was omnipotent, quick to forgive, and ready to protect those who believed (5). He has never questioned God's powers. Until (6) he sees an innocent child hung on the gallows, his father beaten, the massive burial of people who have been tortured and killed.

While living in an environment that is really a cemetery, Elie faces the question of mortality. However, with all hope lost; and (7) his faith crumbling around him, he forces himself to live. Despite everything, he presses on, hoping no one will ever forget the horrors of the Holocaust and pray (8) that this terrible chapter in human history will never happen again.

BEST COPY AVAILABLE

Released Items

English Goal 1

Objective 1.0 — The learner will use strategies and processes that enhance control of communication skills development.

For the seventh underlined item in the essay, choose the correction to be made or "Make no change" if there is no correction.

A lost, and

B lost and

C lost: and

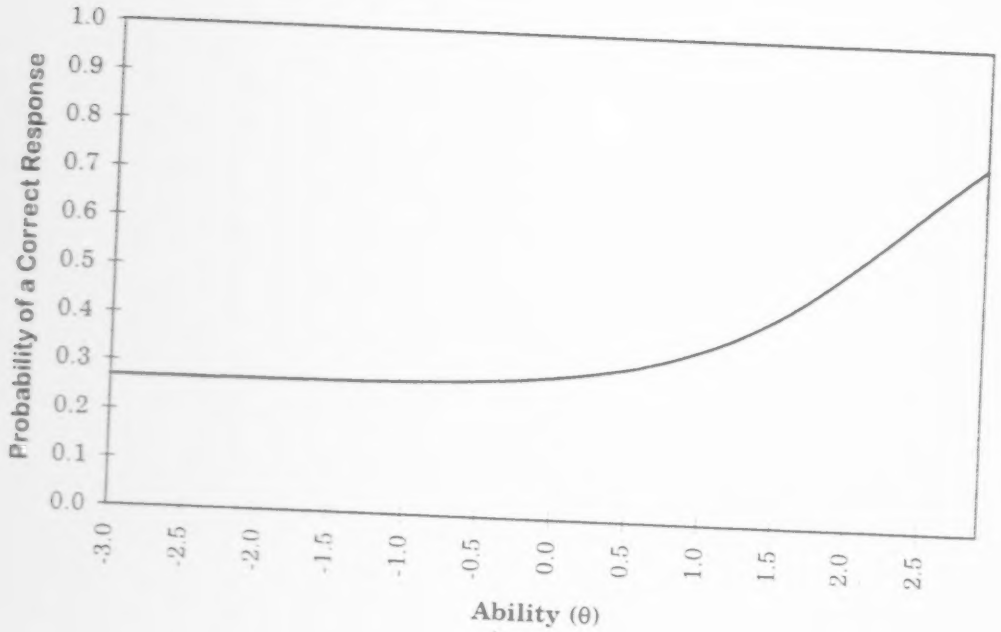
D Make no change.

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2592	7	7	1	2	400	211	17	28

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.32	0.148	0.47	0.776	0.793	2.408	0.821	0.268

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

English Goal 1

Objective 1.0 — The learner will use strategies and processes that enhance control of communication skills development.

For the eighth underlined item in the essay, choose the correction to be made or "Make no change" if there is no correction.

A to pray

B praying

C is praying

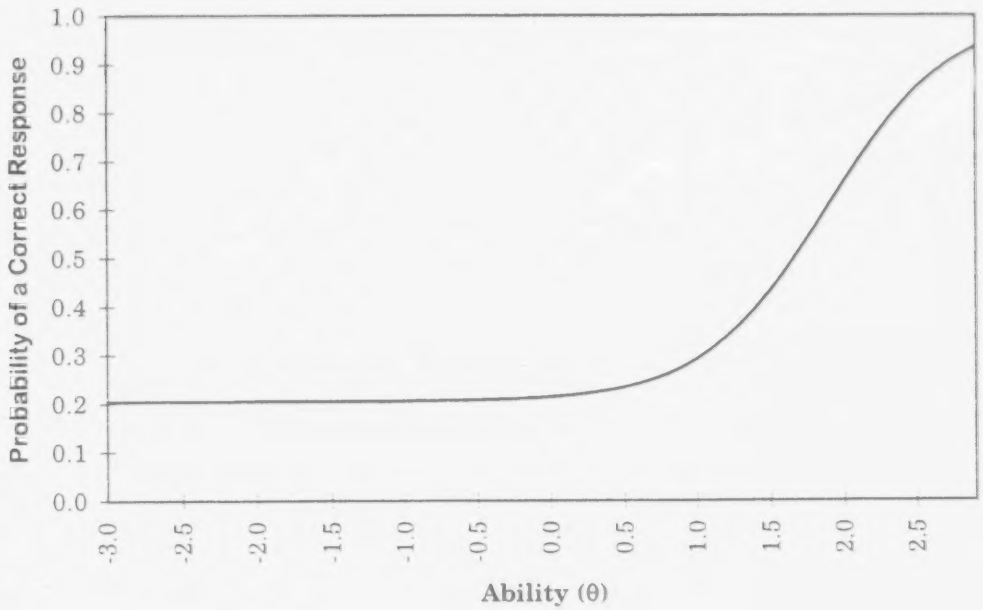
D Make no change.

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2593	7	8	1	2	85	170	119	283

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.26	0.181	0.44	0.489	0.832	1.886	1.402	0.204

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Student Directions: Use the following passage to answer the next two questions.

When I Heard the Learn'd Astronomer

by Walt Whitman

Walt Whitman, who has been called the father of modern poetry, worked as a printer, teacher, and editor before deciding to write full-time. Read the following poem by Walt Whitman and answer the questions.

When I heard the learned astronomer,
When the proofs, the figures, were ranged in columns before me,
When I was shown the charts and diagrams, to add, divide, and measure them,
When I sitting heard the astronomer where he lectured with much applause in
the lecture-room,
How soon unaccountable I became tired and sick,
Till rising and gliding out I wander'd off by myself,
In the mystical moist night-air, and from time to time,
Look'd up in perfect silence at the stars.

From *Leaves of Grass* by Walt Whitman. Public Domain.

Released Items

English I Goal 3

Objective 3.3 — The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.

Which quotation ***best*** relates to the theme of this poem?

A "You can only form the minds of reasoning animals upon Facts;" (*Passage 1*)

B "Does it all add up on the plus side of the ledger?" (*Passage 2*)

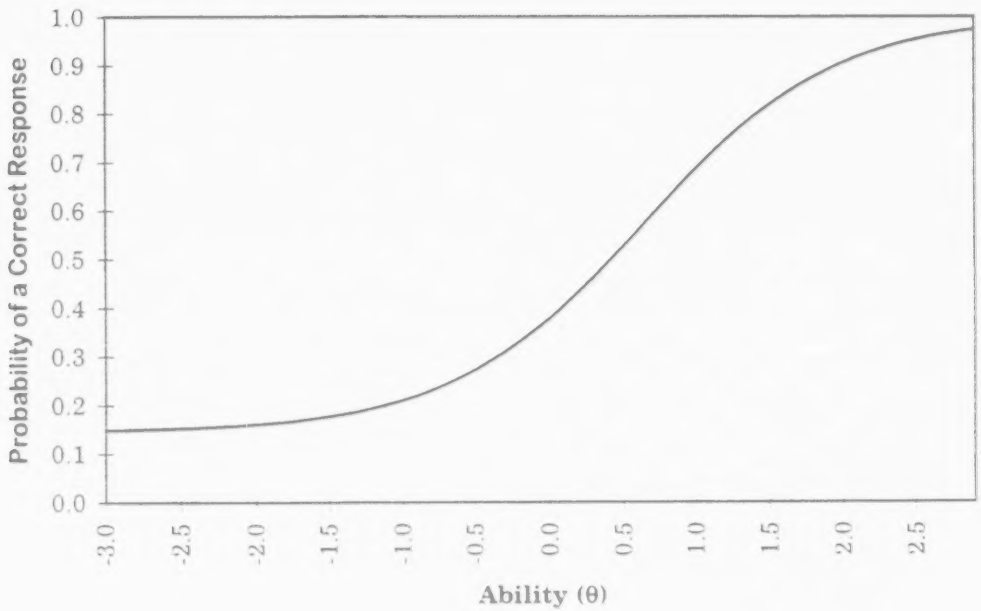
C . . . the principal of Victory School and my eighth grade teacher was special. He liked, really liked every one of us." (*Passage 3*)

D "I did not see that he was also teaching us that education does not come only from between the covers of books, but from the worlds outside the classroom door." (*Passage 4*)

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2715	11	78	3.3	4	90	106	118	282
		-----Bias-----			------IRT Parameters------			
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote	
.47	0.550	0.50	0.913	0.761	0.646	0.900	0.145	

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

English I Goal 2

Objective 2.2 — The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.

This poem is ***best*** classified as which of the following?

A a sonnet

B epic poetry

C lyric poetry

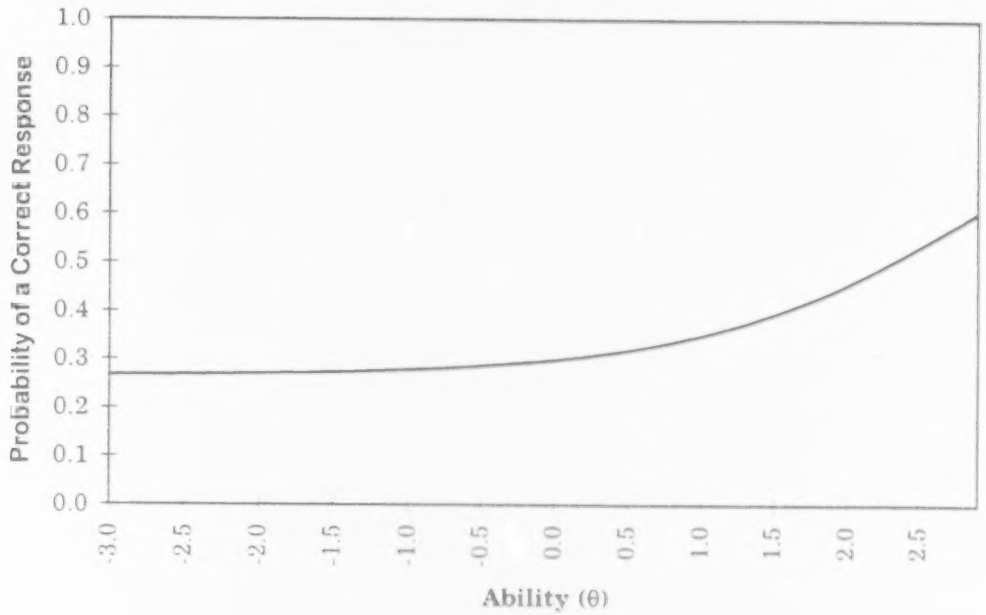
D a ballad

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2716	11	76	2.2	3	145	153	203	96

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asym.ptote
.34	0.130	0.47	1.343	1.090	3.064	0.586	0.265

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

U.S. History Goal 11

Objective 11.2 — Analyze changing relationships between states and the federal government as the role of the federal government continued to expand.

"It is time to check and reverse the growth of government which shows signs of having grown beyond the consent of the governed."

— Ronald Reagan, Inaugural Address, 1981

According to this quotation, President Reagan believed which of the following?

A The federal government had lost control of the citizens it was created to serve.

B The federal government had expanded its role too far into the lives of individual citizens.

C The state governments had taken too much legislative authority from the federal government.

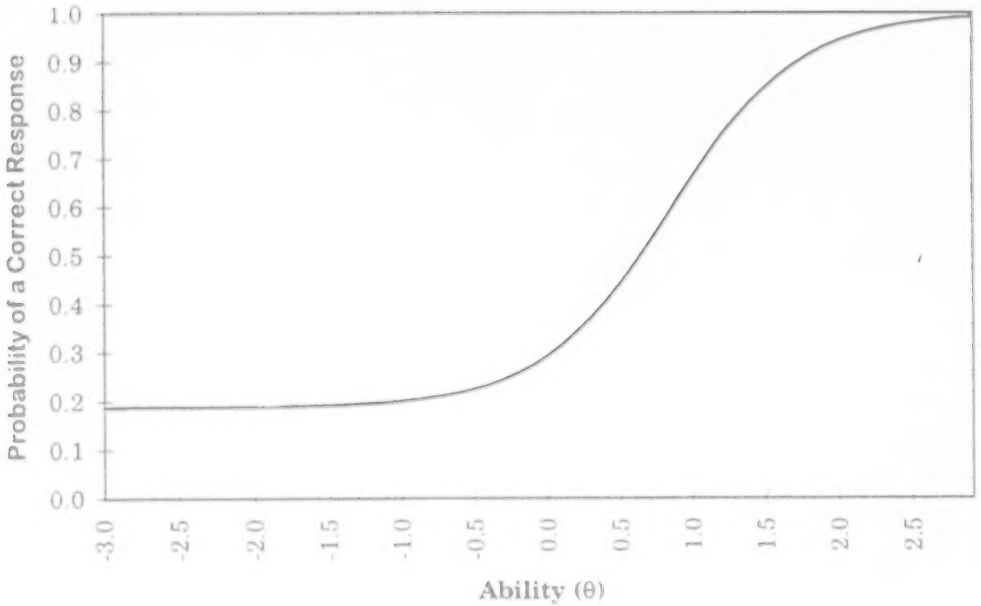
D The state governments were not equipped to deal with the problems of the modern age.

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
3390	2	96	11.2	2	162	324	155	83

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.40	.587	.49	0.999	1.080	0.852	1.321	0.187

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

U.S. History Goal 5

Objective 5.2 — Analyze long-term and immediate causes of the war and assess the extent to which slavery was the primary cause of the conflict.

Frederick Douglass first said, "Slave holders know that the day of their power is over when a Republican President is elected." Later, he said slavery would "be as safe and safer with Lincoln than with a Democrat." Douglass' change of mind is ***best*** explained by which of the following?

A Republican leaders promised not to interfere with existing slavery.

B Republicans planned to enforce the Fugitive Slave Law.

C Republicans pledged to put down slave rebellions.

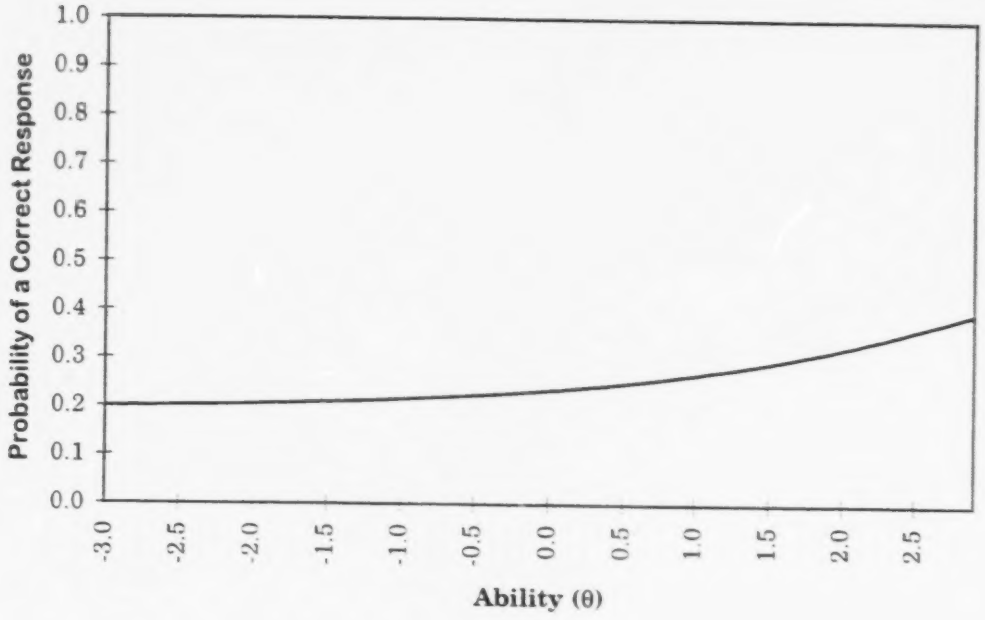
D Republican leaders refused to accept the right of slave states to secede.

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
3331	2	36	5.2	1	248	219	144	195

		-----Bias-----			------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.24	.020	0.43	0.838	0.589	4.634	0.368	0.192

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

U.S. History Goal 6

Objective 6.4 — Evaluate the effects of racial segregation on various regions and segments of American society.

Which of the following is an example of *de facto* racial segregation?

A "Whites Only" signs in restaurants

B housing pattern in large cities

C separate white and black school systems

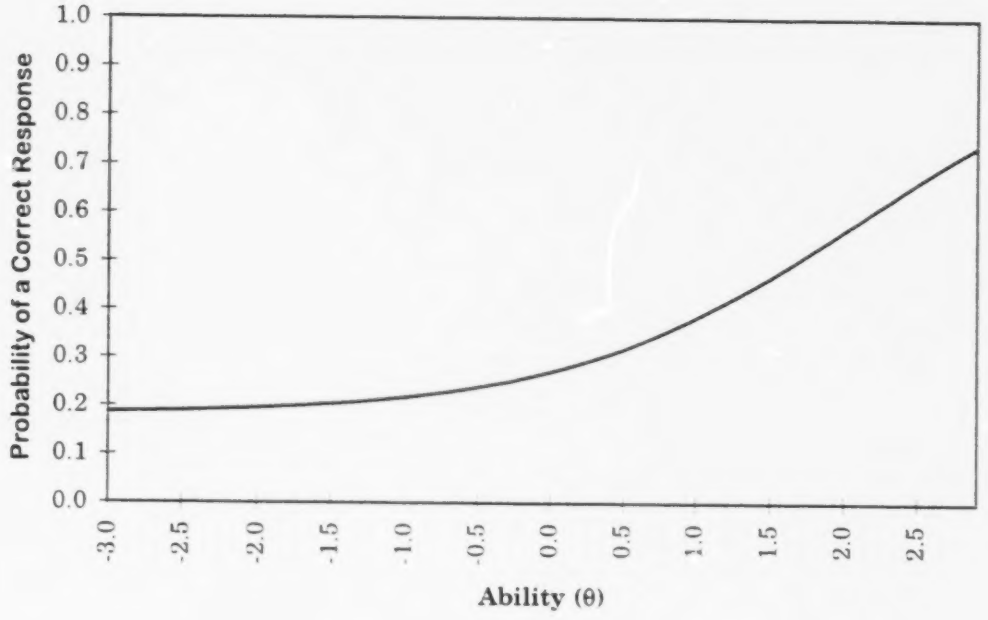
D public water fountains labeled "colored" and "white"

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2882	2	44	6.4	2	178	244	214	162

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.30	.281	.46	0.831	1.003	2.134	0.580	0.181

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Released Items

U.S. History Goal 9

Objective 9.1 — Investigate reasons for the expansion of totalitarian governments during the period.

How did tariffs contribute to the rise of totalitarian dictatorships?

A American arms salespeople were allowed indiscriminately to sell weapons to any country.

B Tariffs disrupted trade and contributed to world economic depression.

C The United States demanded immediate payment of European war debts.

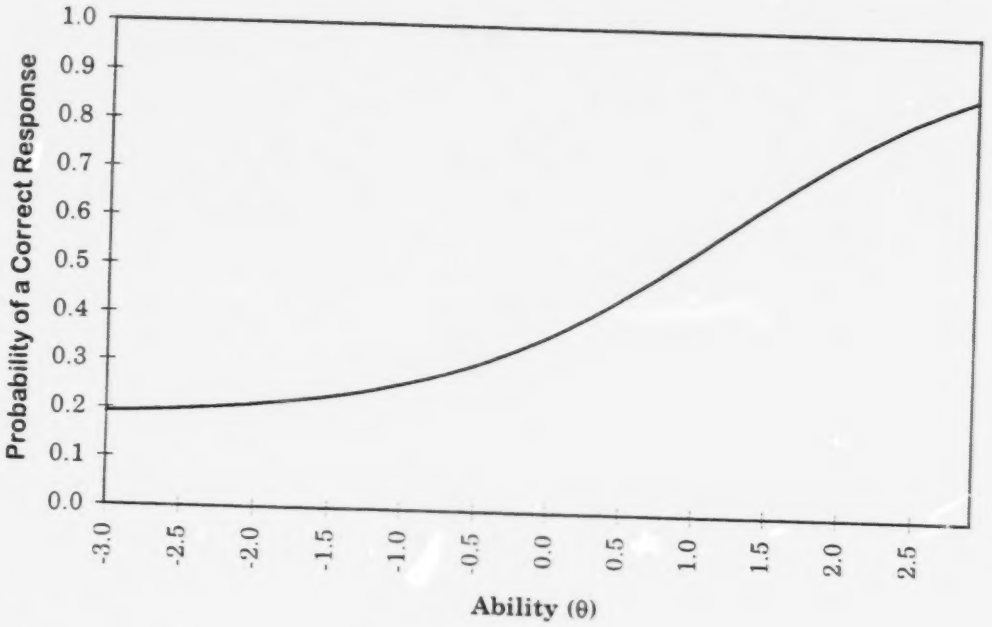
D Foreign-owned businesses in the United States were confiscated by the federal government.

Item Statistics

					--------Choice--------			
Origno	Form	Item	Obj	Key	A	B	C	D
2343	2	69	9.1	2	153	324	166	105

			-----Bias-----		------IRT Parameters------		
P	Bis	Psd	Ethnic	Gender	Threshold	Slope	Asymptote
.40	.362	.49	0.854	0.809	1.233	0.595	0.181

NCTests


Probability of a Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Ability (θ)

Appendix B

Curricula, Test Specifications, and Average Difficulty of Item Pools

Algebra I B–2

Biology B–7

Economic,Legal, and Political Systems B–9

English I B–15

US History B–16

Notes:

Number of Items per Form: This is the typical number of items per form. Some forms may have one more or one less because the objective may be slightly more important than another objective, therefore additional items are needed for curriculum coverage.

Number of Items per Class: This is the number of items that are administered in each class in order to evaluate the implementation of the curriculum.

Difficulty of Pool: This is the average percent correct across all items that measure the goal or objective. (NT = Not Tested)

Algebra I

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will use the language of Algebra.	4	12	0.702
1.1 (c)	Evaluate algebraic expressions.	1	3	0.655
1.2 (c)	Use formulas to solve problems.	1	3	0.662
1.3 (c)	Translate word phrases and sentences into expressions and equations and vice versa.	1	3	0.852
1.4	Use the associative, commutative and distributive properties.	1	3	0.749
2.0	The learner will perform operations with real numbers.	10	30	0.476
2.1 (c)	Simplify real number expressions with and without a calculator.	1	3	0.557
2.2 (c)	Determine the additive or multiplicative inverse of a number.	1	3	0.316
2.3 (c)	Evaluate to determine the absolute value of expressions.	1	3	0.539
2.4 (c)	Raise a real number to an indicated power.	1	3	0.616
2.5 (c)	Write numbers in scientific notation and use this notation with the calculator.	1	3	0.384
2.6 (c)	Distinguish between rational and irrational numbers.	1	3	0.504
2.7 (c)	Find approximations for square roots with and without a calculator.	1	3	0.726
2.8 (c)	Simplify radical expressions.	1	3	0.322
2.9 (c)	Multiply two binomials which contain square roots.	1	3	0.281
2.10 (c)	Compare real number expressions.	1	3	0.539
3.0	The learner will solve equations and inequalities with one variable.	13	39	0.544
3.1	Solve a simple equation by using the addition property of equality and the idea of additive inverse.	1	3	0.778
3.2	Solve a simple equation by using the multiplication property of equality and the idea of multiplicative inverse.	1	3	0.736

Algebra I (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
3.3	Solve an equation graphically and by using more than one property of equality.	1	3	0.740
3.4	Solve an equation which contains similar terms.	1	3	0.748
3.5	Solve an equation which has the variable in both members.	1	3	0.598
3.6	Solve an equation in which the numerical coefficient is a fraction.	1	3	0.515
3.7	Solve a formula for one of its variables or find the value of a variable when values of the other variables are given.	1	3	0.385
3.8 (c)	Use problem solving skills to solve real world, and "word" problems which involve a linear equation or a formula.	1	3	0.371
3.9	Solve a simple equation involving absolute value.	1	3	0.316
3.10	Solve a simple equation containing a radical.	NT	NT	NT
3.11	Find the solution set for a linear inequality when replacement values are given for the variables.	1	3	0.484
3.12	Solve a linear inequality by using transformations.	1	3	0.453
3.13 (c)	Use inequalities to solve problems.	1	3	0.376
3.14	Find the solutions set of combined inequalities.	1	3	0.308
4.0	The learner will demonstrate an elementary understanding of relations and functions.	10	30	0.480
4.1 (c)	Graph and locate sets of real numbers on a number line.	1	3	0.786
4.2	Graph ordered pairs of numbers on the coordinate plane and interpret information related to these sets of points.	1	3	0.820
4.3 (c)	Find the distance between two points on a number line.	1	3	0.483
4.4	Graph a relation on the coordinate plane.	1	3	0.310
4.5	Distinguish between a relation and a function.	1	3	0.313
4.6	Graph a relation given an equation and domain.	1	3	0.309
4.7	Sketch a reasonable graph for a given relationship.	1	3	0.410

Algebra I (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
4.8	Interpret a graph in a real-world setting.	1	3	0.827
4.9 (c)	Use a computer or graphing calculator to explore the graphs of functions.	1	3	0.506
4.10	Compare ordered pairs to the line $y = x$ and interpret the results.	1	3	0.403
5.0	The learner will graph and use linear equations and inequalities.	11	33	0.386
5.1	Determine if data are behaving in a linear fashion.	1	3	0.313
5.2	Find the solution set of open sentences in two variables when given replacement sets for the variables.	1	3	0.486
5.3	Graph a linear equation in two variables.	1	3	0.413
5.4	Graph a line given its slope and y-intercept.	1	3	0.417
5.5	Find the slope of a nonvertical line given the graph of a line or an equation of the line or two points on the line.	1	3	0.365
5.6	Describe the slope in a real world linear relationship using real world terms.	1	3	0.372
5.7	Write the slope-intercept form of an equation of a line.	1	3	0.422
5.8	Write the equation of a line given the slope and one point on the line, or two points on the line.	1	3	0.264
5.9	Write the equation of a line which models a set of real data.	1	3	0.310
5.10 (c)	Use the line which models real data to make predictions.	1	3	0.473
5.11	Graph a linear inequality in two variables.	1	3	0.288
6.0	The learner will graph and solve systems of linear equations and inequalities.	7	21	0.369
6.1	Use a graph to find the solution of a pair of linear equations in two variables.	1	3	0.370
6.2	Graph the solution set of a system of linear inequalities in two variables.	1	3	0.303

Algebra I (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
6.3 (c)	Use a computer or graphics calculator to solve systems of linear equations.	1	3	0.258
6.4	Use the substitution method to find the solution of a pair of linear equations in two variables.	1	3	0.415
6.5	Use the addition or subtraction method to find the solution of a pair of linear equations in two variables.	1	3	0.373
6.6	Use multiplication with the addition or subtraction method to solve systems of linear equations.	1	3	0.376
6.7 (c)	Use systems of linear equations to solve problems.	1	3	0.425
7.0	The learner will perform operations with polynomials.	11	33	0.502
7.1	Add and subtract polynomials.	1	3	0.446
7.2	Multiply monomials.	1	3	0.579
7.3	Find an indicated power of a monomial.	1	3	0.538
7.4	Multiply a polynomial by a monomial.	1	3	0.648
7.5	Find the product of two binomials.	1	3	0.486
7.6	Multiply two polynomials.	1	3	0.447
7.7	Divide two monomials.	1	3	0.525
7.8	Divide a polynomial by a monomial.	1	3	0.484
7.9	Find a common monomial factor in a polynomial.	1	3	0.518
7.10	Factor the difference of two squares.	1	3	0.492
7.11	Factor a simple quadratic trinomial.	1	3	0.461
8.0	The learner will work with ratios, proportions, and percents.	4	12	0.407
8.1	Simplify ratios involving algebraic expressions.	1	3	0.339
8.2 (c)	Solve proportions.	1	3	0.486
8.3 (c)	Use ratios and proportions to solve problems.	1	3	0.425
8.4 (c)	Solve problems involving percents.	1	3	0.398

Algebra I (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
9.0	The learner will explore, graph and interpret nonlinear equations.	11	33	0.266
9.1	Graph a quadratic equation.	1	3	0.305
9.2 (c)	Use an automatic grapher to find the solution to a quadratic equation.	1	3	0.224
9.3	Solve a quadratic equation when one member is in factored form and the other member is zero.	1	3	0.354
9.4	Solve a second degree equation by factoring.	1	3	0.263
9.5 (c)	Use an automatic grapher to relate the solutions of quadratic equations and the *x*-intercepts.	1	3	0.255
9.6 (c)	Understand that the vertex provides the maximum or minimum value of the function.	1	3	0.370
9.7	Solve a quadratic equation in which a perfect square equals a constant.	1	3	0.298
9.8	Solve a quadratic equation by using the quadratic formula.	1	3	0.202
9.9 (c)	Use quadratic equations to solve problems.	1	3	0.275
9.10 (c)	Determine if a set of data represents an exponential function.	1	3	0.186
9.11 (c)	Use formulas, calculators and automatic graphers to explore and solve problems involving exponentials.	1	3	0.226

Biology

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will develop an understanding of the nature of science.	4	11	0.483
2.0	The learner will develop the ability to use science process skills.	50	151	0.439
3.0	The learner will develop the ability to use science manipulative skills.	2	7	0.453
5.0	The learner will have an understanding of the relevance of current topics in biology.	5	14	0.453
CK	The learner will have knowledge of important ideas and concepts.	34	102	0.421
6.0	The learner will understand the chemical and physical basis of life.	30	89	0.461
6.1	Demonstrate knowledge of the differences between living and nonliving things.	3	9	0.478
6.2	Demonstrate knowledge of the cell.	8	24	0.440
6.3	Demonstrate knowledge of the chemical processes of life.	8	23	0.539
6.4	Demonstrate knowledge that living things exist in a state of dynamic equilibrium.	3	9	0.351
6.5	Demonstrate knowledge of anatomy and physiology.	8	24	0.443
7.0	The learner will have an understanding of the continuity of life.	38	114	0.413
7.1	Demonstrate knowledge that living things receive characteristics from the parent organism(s).	8	24	0.417
7.2	Demonstrate knowledge of the chemical basis of heredity.	7	21	0.375
7.3	Demonstrate knowledge that organic variation is important and necessary for species survival.	8	24	0.446
7.4	Demonstrate knowledge of the nature of human origin and development.	3	9	0.404
7.5	Demonstrate knowledge about the diversity and classification of living things.	8	24	0.444

Biology (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
7.6	Demonstrate knowledge of major genetic and environmental factors affecting human health.	4	12	0.390
8.0	The learner will have an understanding of ecology.	28	82	0.401
8.1	Demonstrate knowledge of populations, communities, and ecosystems.	9	27	0.438
8.2	Demonstrate knowledge of the influences of human activity on the environment.	9	26	0.369
8.3	Demonstrate that for all organisms, survival requires suitable responses to the external environment.	10	29	0.426

Economic, Legal, and Political Systems (ELP)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will investigate issues and problems confronting the American economic, legal and political systems.	10	30	0.57
1.1	Describe examples of recurring public problems and issues.	2	5	0.59
1.2	Analyze causes and consequences of recurring social and economic problems and issues.	1	5	0.57
1.3	Evaluate strengths and limitations of the economic, legal and political systems in resolving problems.	2	6	0.50
1.4	Make inferences regarding relationships among economic, legal and political problems.	2	5	0.57
1.5	Explain relationships among local, state, national and international problems.	2	6	0.59
1.6	Compare differing points of view on the proper role of government in the personal lives of citizens.	1	4	0.59
2.0	The learner will explain rights and analyze the obligations of responsible citizenship.	11	33	0.55
2.1	Describe qualifications and procedures for voting and seeking office at the local, state and national levels.	1	4	0.55
2.2	Analyze information on political issues and candidates for political office.	2	5	0.55
2.3	Demonstrate methods of promoting and inhibiting change through political action.	2	5	0.46
2.4	Analyze consequences of compliance or noncompliance with laws governing society.	1	3	0.53
2.5	Explain rights and freedoms available to all citizens of the United States.	2	5	0.55
2.6	Describe situations that benefit from volunteers and display a willingness to volunteer and undertake extra responsibilities for civic welfare at appropriate levels.	1	3	0.58
2.7	Analyze costs and benefits of jury service, voting, seeking office, and civic action in general.	2	5	0.56

ELP (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
2.8	Demonstrate the importance of being a responsible economic decision-maker.	1	3	0.60
3.0	The learner will investigate how and why individuals and groups make economic choices.	10	30	0.55
3.1	Describe basic factors of production (land, labor, capital and entrepreneurial skills) and know their interaction in any economic activity.	2	5	0.56
3.2	Analyze factors which contribute to increased productivity.	2	5	0.56
3.3	Explain why scarcity causes producers and consumers to make choices.	2	5	0.80
3.4	Compare examples of trade-offs and opportunity costs of economic choices.	2	5	0.42
3.5	Analyze the impact of specialization and division of labor on various economic activities.	2	6	0.50
3.6	Explain the impact of investment in capital goods and human capital on the economy.	1	4	0.47
4.0	The learner will analyze features of the economic system of the United States.	11	33	0.54
4.1	Compare characteristics of command, market, traditional, and mixed economies.	1	4	0.48
4.2	Describe examples of how the United States economic system encourages private ownership of property and individual initiative.	1	4	0.51
4.3	Identify characteristics of markets.	1	4	0.50
4.4	Demonstrate how supply, demand and competition affect prices and the availability of goods and services.	1	3	0.59
4.5	Assess the impact on profit of factors such as demand, product quality, management, and training.	1	4	0.61

ELP (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
4.6	Identify distinguishing features of economic institutions (e.g., corporations, banks, stock markets).	1	3	0.52
4.7	Explain the impact of international trade.	1	4	0.51
4.8	Explain the function of money and financial institutions in the United States economy.	1	4	0.62
4.9	Evaluate competing national economic goals and analyze the effects of these goals on various segments of society.	1	2	0.47
5.0	The learner will analyze factors influencing the United States economy.	11	33	0.49
5.1	Analyze the impact of decisions such as saving, using credit, investing, and conserving resources on the economic system.	2	5	0.43
5.2	Make inferences regarding the impact of government regulation on specific economic activities.	1	3	0.47
5.3	Analyze short- and long-term trends in economic activity.	2	5	0.56
5.4	Identify examples of domestic and international economic interdependence.	1	4	0.49
5.5	Analyze short- and long-term effects of taxation and government spending on the United States economy.	1	4	0.51
5.6	Analyze the influence of environmental factors on specific economic activities.	1	3	0.57
5.7	Analyze relationships between economic conditions and political decisions.	2	5	0.43
5.8	Explain how the Gross National Product is used as an indicator of the state of the economy.	1	4	0.47
6.0	The learner will explain the function and importance of the North Carolina and United States Constitutions.	11	33	0.45
6.1	Explain how constitutions define the frame work of governments.	2	5	0.48

ELP (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
6.2	Identify principles found in the United States Constitution which were stated in the Declaration of Independence and explored in the Federalist Papers.	2	6	0.42
6.3	Explain how constitutions grant and limit the authority of public officials and government agencies.	2	5	0.39
6.4	Describe how constitutions may be changed and analyze the impact of specific changes.	1	3	0.53
6.5	Analyze cases which illustrate that the United States Constitution is the supreme law of the land.	2	5	0.45
6.6	Analyze cases which demonstrate how the United States Constitution and Bill of Rights protect the rights of individuals.	1	4	0.46
6.7	Identify modern controversies related to powers of the federal government that are similar to the debates between Federalists and Antifederalists over ratification of the Unites States Constitution.	2	5	0.44
7.0	The learner will explain the structure and functions of local, state and national governments and understand their relationship.	10	30	0.43
7.1	Explain the structure, functions and relationships of the executive, legislative and judicial branches.	2	7	0.41
7.2	Describe services provided by selected government agencies and how they are funded.	2	7	0.46
7.3	Compare major responsibilities of the United States Congress with those of the North Carolina General Assembly and local governing bodies.	2	6	0.40
7.4	Analyze costs and benefits of governmental functions.	1	4	0.46
7.5	Identify key government officials, how they are chosen, and their duties and responsibilities.	2	6	0.42

ELP (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
8.0	The learner will explain why laws are needed and how they are enacted, implemented and enforced.	11	33	0.51
8.1	Illustrate the need for law in society.	1	4	0.59
8.2	Compare examples of criminal, civil, constitutional and administrative law.	2	5	0.55
8.3	Analyze stages involved in the legislative process.	2	5	0.44
8.4	Analyze methods used to inform citizens of the purpose and meaning of laws.	2	5	0.50
8.5	Describe the role of the executive branch in the implementation of laws and policies.	1	4	0.51
8.6	Compare responsibilities, jurisdictions and methods of individual law enforcement agencies.	2	5	0.50
8.7	Evaluate various methods used by society to deal with criminal and antisocial behavior.	2	5	0.51
9.0	The learner will explain how the political and legal systems provide for balancing competing interests and resolving conflicts.	10	30	0.45
9.1	Evaluate the role of debate, consensus, compromise and negotiation when resolving conflicts.	2	6	0.51
9.2	Describe the adversarial nature of the judicial processes.	2	5	0.55
9.3	Evaluate the role of debate and compromise in the legislative process.	2	5	0.39
9.4	Identify the jurisdiction of state and federal courts.	2	5	0.45
9.5	Analyze roles of individual citizens, political parties, the media, and other interest groups in public policy decisions.	2	6	0.35
9.6	Explain how local government agencies balance interests and resolve conflicts.	1	3	0.39

ELP (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
10.0	The learner will evaluate the influence of ethical and moral principles and religious beliefs on the development of our economic, legal and political systems.	5	15	0.47
10.1	Analyze individual and group decisions are made on the basis of a variety of standards including aesthetic, pragmatic and ethical.	2	5	0.44
10.2	Analyze examples of conduct by public officials, corporate officers and private citizens in a variety of situations and evaluate their conduct in terms of given criteria.	1	3	0.49
10.3	Evaluate positions on a variety of issues against given criteria.	1	3	0.45
10.4	Analyze the meaning and the influence on our society of the religion clauses of the First Amendment to the United States Constitution.	1	4	0.52

English I

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will use strategies and processes that enhance control of communication skills development.	31	93	0.551
2.0	The learner will use language for the acquisition, interpretation, and application of information.	26	79	0.462
2.1	The learner will identify, collect, or select information and ideas.	2	4	0.530
2.2	The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.	22	67	0.465
2.3	The learner will apply, extend, and expand on information and concepts.	2	8	0.440
3.0	The learner will use language for critical analysis and evaluation.	15	45	0.473
3.1	The learner will assess the validity and accuracy of information and ideas.	0	0	0.525
3.2	The learner will determine the value of infromation and ideas.	2	6	0.460
3.3	The learner will develop criteria and evaluate the quality, relevance, and importance of the information and idcas.	13	39	0.474

U.S. History

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will analyze those elements in the American colonial experience that led to separation from England.	8	24	0.440
1.1	Describe how geographic diversity influenced economic, social, and political life in colonial North America.	2	6	0.441
1.2	Describe the contributions of various racial, ethnic, and religious groups including African Americans and Native Americans to the development of a new culture.	2	6	0.538
1.3	Elaborate on the sources of American nationalism.	1	3	0.340
1.4	Distinguish between immediate and long-term causes of the American Revolution.	1	3	0.387
1.5	Assess the importance of military engagements, personalities, and geopolitical factors in the defeat of the British.	2	6	0.415
2.0	The learner will apply ideas of self-government as expressed in America's founding documents.	9	27	0.480
2.1	Trace the development of concepts of self-government in British North America from the Mayflower Compact to the Declaration of Independence.	2	6	0.456
2.2	Associate ideas in the founding documents with their European origins.	2	6	0.385
2.3	Analyze the Declaration of Independence and the Constitution of the United States as expressions of self-government.	2	6	0.497
2.4	Evaluate the arguments of the Federalist and the Antifederalist papers as expressions of differing theories about self-government.	1	3	0.445
2.5	Judge the extent to which the Bill of Rights extended the Constitution.	2	6	0.566

U.S. History (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
3.0	The learner will judge the effectiveness of the institutions of the new nation in completing its independence. (1781-1815)	6	18	0.400
3.1	Identify major domestic problems of the nation under the Articles of Confederation and judge the extent to which they were resolved by the new Constitution.	1	3	0.366
3.2	Judge the extent to which the institutions of the new nation protected the liberties of all its inhabitants.	1	3	0.632
3.3	Trace the development of religious liberty and toleration in the new nation.	1	3	0.439
3.4	Analyze the effects of territorial expansion and the admission of new states to the Union.	1	3	0.388
3.5	Assess commercial and diplomatic relationships with Britain, France, and other nations.	1	3	0.354
3.6	Evaluate the extent to which the United States was "a nation at risk" until 1815.	1	3	0.395
4.0	The learner will assess the contending forces of nationalism and sectionalism in the period 1815 - 1850.	8	24	0.400
4.1	Map westward expansion and make inferences about its importance to African Americans and Native Americans.	2	6	0.428
4.2	Analyze economic developments and judge their effects on nationalism and sectionalism.	2	6	0.390
4.3	Assess political events and personalities in terms of their influence on nationalistic or sectional trends.	1	3	0.375
4.4	Analyze literary and artistic movements of the period as contributors to nationalism and sectionalism.	2	6	0.386
4.5	Evaluate the role of religion in the debate over slavery and in other social movements and changes of the period.	1	3	0.434

U.S. History (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
5.0	The learner will evaluate the Civil War and Reconstruction as an affirmation of the power of the national government.	7	21	0.430
5.1	Elaborate on economic, social, and political conditions in the decade preceding the Civil War.	2	6	0.457
5.2	Analyze long-term and immediate causes of the war and assess the extent to which slavery was the primary cause of the conflict.	2	6	0.392
5.3	Trace important military and political events of the war period, and judge their significance to the outcome of the conflict.	1	3	0.454
5.4	Judge immediate and long-term effects of Reconstruction on the daily lives of people as well as on the politics and economy of the former Confederate states.	2	6	0.428
6.0	The learner will interpret economic, social, and political trends of the late nineteenth and early twentieth centuries.	6	18	0.400
6.1	Describe innovations in technology and business practices and assess their impact on the economy.	1	3	0.421
6.2	Make inferences about the influence of immigration and rapid industrialization on urban life.	1	3	0.530
6.3	Trace the development of labor unions and judge their effects on economic arrangements and the lives of working people.	1	3	0.354
6.4	Evaluate the effects of racial segregation on various regions and segments of American society.	1	3	0.379
6.5	Trace the rise and decline of Populism and Progressivism and judge their effectiveness as economic, social, and political movements.	1	3	0.370
6.6	Analyze the influence of growing religious pluralism on American society.	1	3	0.367

U.S. History (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
7.0	The learner will analyze the causes and effects of United States involvement in international affairs.	10	30	0.390
7.1	Trace the emergence of the United States as an increasingly significant international power in the late 19th and early 20th centuries.	3	9	0.385
7.2	Analyze the causes of United States involvement in World War I, and assess the effects of the war on the United States and other nations.	4	12	0.395
7.3	Assess the significance of the war experience on United States foreign and domestic policies of the 1920's and 1930's.	3	9	0.393
8.0	The learner will appraise the economic, social, and political changes of the decades of the 'Twenties and 'Thirties.	8	24	0.400
8.1	Elaborate on the cycle of economic boom and bust in the 'twenties and 'thirties and analyze the extent of "prosperity" for different segments of society.	2	6	0.433
8.2	Make inferences about social, intellectual, and technological change based on an analysis of lifestyles of the period.	2	6	0.406
8.3	Describe challenges to tradition in religion, race, and gender during the period.	2	6	0.404
8.4	Assess the impact of New Deal reforms in enlarging the role of the federal government in American life.	2	6	0.368
9.0	The learner will analyze and evaluate the significance of causes, events, and effects of the World War II Era.	10	30	0.370
9.1	Investigate reasons for the expansion of totalitarian governments during the period.	2	6	0.343
9.2	Trace the course of events that resulted in a new outbreak of worldwide war and analyze the role of the United States in those events.	3	9	0.369

U.S. History (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
9.3	Identify major campaigns and personalities from the World War II era, and assess their importance to the conduct of the war.	2	6	0.367
9.4	Describe and analyze the effects of the war on American economic, social, and political life.	3	9	0.392
10.0	The learner will trace economic and social developments and assess their significance for the lives of Americans in the last half century.	10	30	0.400
10.1	Identify technical innovations that have significantly affected American life and judge the importance of their influence on our behavior.	2	6	0.428
10.2	Elaborate on the suburbanization of American society and make inferences about its importance to our economic and social institutions.	2	6	0.373
10.3	Trace major events of the Civil Rights Movement and evaluate the impact of the movement on institutions and the lives of citizens.	2	6	0.433
10.4	Assess the importance of growing religious pluralism and racial and ethnic diversity in American society.	2	6	0.433
10.5	Analyze the course of the United States economy in the last half century.	2	6	0.338
11.0	The learner will analyze changes in American political life in the last half-century.	10	30	0.370
11.1	Trace changes in political party alignment and voter behavior.	3	9	0.389
11.2	Analyze changing relationships between states and the federal government as the role of the federal government continued to expand.	2	6	0.290
11.3	Analyze relationships and actions of the three branches of the federal government in terms of their influence on the lives of citizens.	3	9	0.378

U.S. History (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
11.4	Assess the influence of phenomena such as television on the conduct of American politics.	2	6	0.403
12.0	The learner will evaluate the conduct of United States foreign policy over the last half-century.	8	24	0.360
12.1	Trace the course of the "cold war" and judge its impact on American society.	2	6	0.334
12.2	Elaborate on changes in the direction of foreign policy toward various world areas over the period.	2	6	0.343
12.3	Examine the role of organizations established to maintain peace and judge their continuing effectiveness.	2	6	0.404
12.4	Identify causes of United States involvement in foreign wars since World War II and judge the influence of our involvement on American society.	2	6	0.363

Appendix C

North Carolina Technical Advisory Group

Dr. John Fremer (Chairperson)
Educational Testing Service

Dr. Mark Applebaum
Vanderbilt University

Dr. Lloyd Bond
University of North Carolina at Greensboro

Dr. Joy McLarty
American College Testing

Dr. David Thissen
L.L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Dr. Gary Williamson
Educational Consultant, North Carolina

Appendix D

Sample Item with Development, Review, and Psychometric Information

Item Specifications Sheet D–2

Item Review Sheet D–3

Item Review Summary Report By Item D–4

Item Record with Field Test Information D–5

ITEM SPECIFICATIONS SHEET

02474

CURRICULUM OBJECTIVE:

9.10 Determine if a set of data represents an exponential function.

CURRICULUM SOURCE:

DIFFICULTY LEVEL: 1 = EASY, 2 = MEDIUM, (3) = HARD	THINKING SKILLS: LOWER (1), HIGHER (2)	PAGEMAKER: 1 = YES, 0 = NO
ITEM WRITER NUMBER: 28		ARTWORK: 1 = YES, 0 = NO

ALGEBRA I TEST ITEM *(FINAL DRAFT)*

The frequencies (cycles per second) of the sound of A's on a piano increase in an exponential fashion as you move up the keyboard. Which of the following are frequencies of Key A?

A) 55, 110, 440, 3520, 56320

B) 55, 110, 220, 440, 880

C) 55, 110, 165, 220, 275

D) 55, 105, 155, 205, 255

CORRECT ANSWER B

EDIT cm

Did You...

1. focus directly on the objective?
2. write stem as a complete statement of question?
3. write foils of equal length with *only* one correct answer?
4. use same context and similar ideas in foils?
5. avoid using negatives in the foils?
6. arrange continuous foils in logical order?
7. make each foil credible?
8. check punctuation, spelling, and grammatical structure of item?
9. use artwork *only* when necessary?
10. practice fair representation in sex and race, avoiding culture specific references?

Numbers for Keypunch only	
1,2	I.D.
3,4	Goal
5,6	Obj
7	Think
8	Diff
9	Diagram
10-13	Item No.

Algebra I Item Review

9.10 Determine if a set of data represents an exponential function.

Difficulty Level: Hard

2474

00. The frequencies (cycles per second) of the sound of A's on a piano increase in an exponential fashion as you move up the keyboard. Which of the following are frequencies of key A?

A 55, 110, 440, 3520, 56320

B 55, 110, 220, 440, 880

C 55, 110, 165, 220, 275

D 55, 105, 155, 205, 255

14 **Correct Answer:** _____

15 **Thinking Skill Level:**
- ☐ Knowledge (1)
- ☐ Organizing (4)
- ☐ Applying (5)
- ☐ Analyzing (6)
- ☐ Generating (7)
- ☐ Integrating (8)
- ☐ Evaluating (9)

	CONCEPTUAL			LANGUAGE			FORMAT			DIAGRAM		
	Yes	Marginal	No	Yes	Marginal	No	Yes	Marginal	No	Yes	Marginal	No
	16 ☐	☐	☐	26 ☐	☐	☐	33 ☐	☐	☐	40 ☐	☐	☐
17	☐ Objective match			27 ☐ Appropriate for age			34 ☐ Logical order of foils			41 ☐ Necessary		
18	☐ Fair representation			28 ☐ Punctuation, spelling, grammar			35 ☐ Familiar presentation style			42 ☐ Clear		
19	☐ No cultural bias			29 ☐ No excess words			36 ☐ Print size and type			43 ☐ Relevant		
20	☐ Clear statement			30 ☐ No stem/foil clues			37 ☐ Mechanics and appearance			44 ☐ Unbiased		
21	☐ Single problem			31 ☐ No negatives in foils			38 ☐ Equal length foils			45 ☐ Other		
22	☐ One best answer			32 ☐ Other			39 ☐ Other					
23	☐ Common context in foils											
24	☐ Each foil credible											
25	☐ Other											

46 **OVERALL RATING:** ☑ Acceptable ☐ Acceptable with modifications ☐ Discard Item

BEST COPY AVAILABLE

12/18/92

Report By Item
ALGEBRA I ITEM REVIEW
Item number: 2474

Goal: 9.10

THINKING SKILL LEVEL		ANSWER	
Knowledge:	0	A:	1
Organizing:	0	B:	2
Applying:	1	C:	0
Analyzing:	1	D:	0
Generating:	1	E:	0
Integrating:	0		
Evaluating:	0		

		No	Marginal	Yes
Conceptual	:	0	1	2
Fair Representation	:	1		
Language	:	0	0	3
Format	:	0	0	3
Diagram	:	0	0	0

OVERALL RATING

Acceptable :	3
Modify :	0
Discard :	0

TOTAL REVIEWERS: 3

COMMENTS
18 64 => PERSONS WITH MUSICAL BACKGROUND COULD HAVE AN ADVANTAGE

Algebra I

Objective: 9.10 Determine if a set of data represents an exponential function.

On Grade

Algebra I Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
2474	4	87	9.10	2	85	224	180	104

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Threshold	IRT Slope	IRT Asymptote
.31	0.160	.46	0.723	0.947	2.796	0.455	.211

Above Grade

87. The frequencies (cycles per second) of the sound of A's on a piano increase in an exponential fashion as you move up the keyboard. Which of the following are frequencies of key A?

A 55, 110, 440, 3520, 56320

B 55, 110, 220, 440, 880

C 55, 110, 165, 220, 275

D 55, 105, 155, 205, 255

4 87
WEAK PRED

Psychometric Approval		Curriculum Approval	
Yes	No	(Yes)	No

Appendix E

Excerpt from

"Item Response Theory for Scores on Tests including Polychotomous Items with Ordered Responses"

by David Thissen, Mary Pommerich, Kathleen Billeaud, and Valerie S.L. Williams
L.L. Thurstone Psychometric Laboratory of the University of North Carolina at Chapel Hill.
Research Report Number 94-2, Published May, 1994.

Item Response Theory for Scores on Tests including Polychotomous Items with Ordered Responses

David Thissen
Mary Pommerich
Kathleen Billeaud
Valerie S.L. Williams
L.L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Abstract

Item response theory (IRT) provides a ready mechanism for scoring tests including any combination of rated constructed-response and keyed multiple-choice items, in that each response pattern is associated with some modal or expected *a posteriori* estimate of proficiency. However, various considerations that frequently arise in large-scale testing make response-pattern scoring an undesirable solution. In this paper, we describe methods based on IRT that provide scaled scores, or estimates of proficiency, for each summed score for rated responses, or for combinations of rated responses and multiple-choice items. These methods may be used to combine the useful scale properties of IRT-based scores with the practical virtues of a scale based on the summed score for each examinee.

The research reported here was supported by the North Carolina Department of Public Instruction, in conjunction with the development of the North Carolina End of Grade Testing Program. We thank Richard Luecht, Robert McKinley, Robert Mislevy, James Ramsay, and Linda Wightman for their help in the course of this work.

One feature of item response theory (IRT) is that it provides a score scale that is more useful than the summed score, percentage-correct, or percentile scales for many purposes, e.g., for construction of developmental scales or for calibration of tests comprising different types of items or exercises. With the exception of the so-called Rasch-family of models for which the summed score is a sufficient statistic for the characterization of the latent variable (θ) (Rasch, 1960; Masters and Wright, 1984), under IRT models each response pattern is associated with a unique estimate of θ. These estimates of θ may be used as scaled response pattern scores; they have the advantage that they extract all information available in the item responses. In addition, the IRT model produces estimates of the probability that each response pattern will be observed in a sample from a specified population.

However, in applied measurement contexts, it is often desirable for various reasons to consider the implications of the IRT analysis for summed scores, rather than response patterns, even if the IRT model used is not part of the Rasch family. For various practical reasons it may be desirable to report IRT scaled scores based on the summed score, rather than the scaled scores that are associated with each response pattern. In addition, it may be useful to compute model-based estimates of the summed score distribution, e.g., to create percentile tables for use as an interpretive aid for score reporting. Model-based estimates of the summed score distribution may also have value as a statistical diagnostic of the goodness of fit of the IRT model, including the validity of the assumed underlying population distribution.

Many contemporary tests include extended constructed-response items, for which the item scores are ordered categorical ratings provided by judges. In some cases, the constructed-response items comprise the entire test; in other cases, there are multiple-choice items as well. In either case, some total score is often required, combining the judged ratings of the constructed-response items (and the item scores on the multiple-choice items, if any are present). Simple summed scores may not be very useful in this context, because of the problems associated with the selection of relative weights for the different items and item types, and because the constructed-response items are often on forms of widely varying difficulty.

Item Response Theory for Summed Scores

For any IRT model for items indexed by i with ordered item scores $k = 0, \ldots K_i$, the likelihood for any summed score $j = \sum k_i$ is

$$L_j(\theta) = \sum_{\in\, j = \sum k_i}^{patterns} L(\mathbf{k}|\theta) \ ,$$

where the summation is over the response patterns with total score j. The likelihood for each response pattern is

$$L(\mathbf{k}|\theta) = \prod_i T_{k_i}(\theta)\phi(\theta) \ ,$$

where $T_{k_i}(\theta)$ is the trace line for category k of item i—the conditional probability of response k to item i given θ—and $\phi(\theta)$ is the population density. Thus, the likelihood for each score is

$$L_j(\theta) = \sum_{\in\, j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta)\phi(\theta) \ .$$

and so the probability of each score j is

$$P_j = \int L_j(\theta)\, d\theta \ ,$$

or

$$P_j = \int \sum_{\in\, j = \sum k_i}^{patterns} L(\mathbf{k}|\theta)\, d\theta \ ,$$

or, most intimidatingly,

$$P_j = \int \sum_{\epsilon\, j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta)\phi(\theta)\, d\theta \,. \qquad (1)$$

Given an algorithm to compute the integrand in Equation 1, it is straightforward to compute the average (or expected *a posteriori*, or *EAP*) scaled score (Bock and Mislevy, 1982) associated with each score,

$$EAP(\theta | j = \sum k_i) = \frac{\int \theta \; L_j(\theta) d\theta}{P_j} \,, \qquad (2)$$

and the corresponding standard deviation,

$$SD(\theta | j = \sum k_i) = \sqrt{\frac{\int [\theta - EAP(\theta | \sum k_i)]^2 \; L_j(\theta)\, d\theta}{P_j}} \,. \qquad (3)$$

The values computed using Equation 2 may be tabulated and used as the IRT scaled-score transformation of the raw scores, and the values of Equation 3 may be used as a standard description of the uncertainty associated with those scaled scores.

The score-histogram created using the values of Equation 1 may be used to construct summed-score percentile tables; if the IRT model fits the data, this can be done accurately using only the item parameters, for any group with a known population density. Thus, percentile tables for summed scores can be constructed using item tryout data, before the operational test is administered. This same histogram may also prove useful as a diagnostic statistic for the goodness of fit of the model, by comparing the modeled representation of the score distribution to the observed data.

Algorithms for Computing $L_j(\theta)$

Lord (1953) used heuristic procedures to describe the difference between the distribution of summed scores, $L_j(\theta)$, and the underlying distribution of θ, $\phi(\theta)$ [see also Lord and Novick (1968, pp. 387-392)]. However, practical calculation of the summed score distribution implied by an IRT model has awaited both contemporary computational power and solutions to the apparently intractable computational problem.

Brute force evaluation of Equation 1, requiring the computation of $\prod(K_i + 1)$ likelihoods, is easy for a few items; but it is inconceivable for many items. Brute force may be extended to moderate numbers of items (i.e., about 20) by using an algorithm involving the computation of each pattern likelihood from some other previously-computed pattern likelihood by a single (list) multiplication; this approach is used in the computer program TESTFACT (Wilson, Wood, & Gibbons, 1991). For binary items, by carefully ordering the computation of the likelihoods for the 2^I patterns (where I is the number of items), such an algorithm can compute all 2^I likelihoods at a computational cost of only a single (list) multiplication for each (Thissen, Pommerich, and Williams, 1993). Nevertheless, due to the exponential computational complexity of the brute force approach, this algorithm cannot be extended to more items regardless of improvements in computational speed.

Lord and Novick (1968, p. 525) stated that "approximations appear inevitable," and suggested the use of an approximation to the compound binomial, attributed to Walsh (1963), to compute the likelihood of a summed score for binary items as a function of θ. For I items, this Taylor-series expansion has I terms; however, in practice the first two terms suffice for acceptable accuracy. The two-term version of the approximation is:

$$\sum_{\epsilon\, j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta) \equiv p_I(j) + \frac{I}{2} VC(j) \,, \qquad (4)$$

where

$$p_I(j) = \begin{cases} \binom{I}{j} M^j (1-M)^{I-j} & \text{for } j = 0, 1, \ldots, I; \\ 0 & \text{otherwise} \end{cases} ,$$

$$C(j) = \sum_{v=0}^{2} (-1)^{v+1} \binom{2}{v} p_{I-2}(j-v) ,$$

$$V = \frac{1}{I} \sum_i [T_{1_i}(\theta) - M]^2 ,$$

and

$$M = \frac{1}{I} \sum_i T_{1_i}(\theta) .$$

Yen (1984) used this approximation to develop an algorithm to compute the mode of

$$\sum_{\epsilon \; j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta)$$

for use as a scaled score for examinees with summed score j on a binary test using the three-parameter logistic model. She reported that the two-term Taylor expansion produced noticeably better results than the one-term solution, which is simply an inverse transformation of the test characteristic curve; but the three- and four-term solutions appeared not to add useful precision.

The approximation in Equation 4 may also be substituted for the sum of products in Equations 1, 2, and 3 to compute P_j, $EAP(\theta|j = \sum k_i)$, and $SD(\theta|j = \sum k_i)$. When the results for the two-term approximation were compared to the correct (brute force) results for one of the 20-item examples used by Yen (1984), the error of approximation was usually less than 0.001 for $EAP(\theta|j = \sum k_i)$, and $SD(\theta|j = \sum k_i)$ (Thissen, et al., 1993). Exceptions tended to be the perfect scores, for which the second term of the two-term approximation is zero; there, the approximation could be off by as much as 0.05. The error of approximation for P_j tended to be of the order of 0.0001. For practical use in constructing score-reporting tables, which usually use no greater precision than tenths of a standard deviation for the scores and their standard errors, and integral values for percentile tables, this degree of precision appears to be sufficient. However, the approximation in Equation 4 is still somewhat computationally burdensome, and no generalization has been offered for items with more than two response categories.

The problem of the computational burden is solved by an alternative procedure briefly described by Lord and Wingersky (1984). Abandoning the contention of Lord and Novick (1968, p. 525) that "approximation is inevitable," Lord and Wingersky described a simple recursive algorithm for the computation of

$$L_j^I(\theta) = \sum_{\epsilon \; j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta)$$

for binary items. The algorithm is based on the distributive law, and generalizes readily to items with any number of response categories.

The generalization follows, using the notation $i = 0, 1, \ldots I$ for the items, $k = 0, 1, \ldots K_i$ for the response categories for item i, and $T_{k_i}(\theta)$ for the trace line for category k of item i. In addition, the summed scores for a set of items $[0 \ldots I^*]$ are $j = 0, 1, \ldots \sum_{I^*} K_i$ and the likelihood for summed score j for a set of items $[0 \ldots I^*]$ is $L_j^{I^*}(\theta)$; the population distribution is $\phi(\theta)$.

The generalized recursive algorithm is:

Set $I^* = 0$

$L_j^{I^*}(\theta) = T_{j_{I^*}}(\theta)$, for $j = 0, 1, \ldots K_{I^*}$

Repeat:

For item $I^* + 1$ and scores $j = 0, 1, \ldots \sum_{I^*} K_i$

$$L_{j+k}^{I^*+1}(\theta) = \sum_{k_{I^*+1}} L_j^{I^*}(\theta) \, T_{k_{I^*+1}}(\theta)$$

Set $I^* = I^*+1$

Until $I^* = I$.

For a sample from a population with distribution $\phi(\theta)$, the likelihood for score j is

$$L_j(\theta) = L_j^I(\theta)\,\phi(\theta)$$

and $EAP(\theta|j = \sum k_i)$, $SD(\theta|j = \sum k_i)$, and $P_j(\theta)$ can be computed by integrating $L_j(\theta)$.

No particular parametric form for the trace lines is assumed in the formulation of the recursive algorithm. We have used the three-parameter logistic in work with binary-scored multiple-choice items, and Samejima's (1969) graded model for multiple-category rated items. However, in principle, any trace lines could be used, such as the nonparametric kernel smooths described by Ramsay (1991). The algorithm would produce perfectly accurate, if silly, results if it were used with items for which the responses are *not* ordered. The results would be silly because the response patterns included in any particular summed score would not tend to have likelihoods concentrated near the same values of θ, and so such summed-score likelihoods would tend to be very flat with very large standard deviations.

Nevertheless, the algorithm is completely general. An implementation for the LISP-STAT computing environment (Tierney, 1990) is given in the appendix.

Appendix F

Bias Review Materials

Item Bias Review Information and Directions F–2

Item Bias Review Sheet: Rejected Items F–4

Item Bias Review

Background

To develop achievement tests that are valid, reliable, and educationally appropriate, the North Carolina Department of Public Instruction carries out a series of operations that take several years. In a broad overview, the procedures involve curriculum definition, test design, item writing, item editing, item review, field testing, analyses of field test data, further item editing and review, selection of items for tests, review of tests "as tests", final editing, and then test administration. All item reviews are accomplished by North Carolina teachers and other professional educators. One of the purposes of the item review is to ensure that the test items do not reflect any cultural bias or stereotyping.

All test items have already been reviewed for cultural bias. However, at this time statistical analyses have been performed on the field test data to detect gender and ethnic bias. Items are flagged as "biased" by the statistical techniques if an identified group (males or females, blacks or whites) performs better than would be expected from their overall proficiency in the content area measured by the test. These items require a closer look: judgments must be made about whether the difference in performance on the item is relevant to what the test is intending to measure. In other words, is what is measured by the test item, and the context in which it is measured, something that should be taught as part of the curriculum? If not, then the item is biased and should not be used on the tests. It is these judgments which will determine if an item is eliminated from the item pool due to bias.

Instructions

Enclosed are test items that have been flagged as potentially biased. Each test item is on a separate sheet. Passages and/or supporting materials are included for some questions. At the top of the sheet is the curricular objective the item is intending to measure, followed by statistics reflecting item performance on the field test. Of particular interest to your item review are the following:

- **Origno** - a five-digit item number that uniquely identifies the item
- **P** - the proportion of test-takers that got the item correct
- **Key** - the correct answer choice (1=A, 2=B, 3=C, 4=D)
- **Ethnic/Gender Bias**
  –if greater than 1.5, then the item favors females or whites
  –if less than .66, then the item favors males or blacks
  (Note that the group favored by the item is written in the bottom left-hand corner of the page. If an item favors females, then it is biased against males, etc.)
- **Choice** - the number of test-takers selecting each answer choice

1

The other numbers on the form reflect other characteristics of the item and can be ignored for this purpose. If the item was tested in the next grade level as well, statistics for that grade level are presented in the "Above Grade" box.

When reviewing the test items, keep in mind the following five questions:

1. Does the item contain any offensive gender, ethnic, and/or regional content?
2. Does the item contain gender, ethnic, or cultural stereotyping?
3. Does the item contain activities that will be more familiar to one group than another?
4. Do the words in the item have a different meaning in one group than in another?
5. Could there be group differences in performance that are *unrelated* to proficiency in the content area?

If your answer is *Yes* to any of the five questions, record the five-digit "Origno" and check the appropriate column(s) on the Item Bias Review Sheet. You should comment on all rejected items on your copy of the item sheets. **If the item is acceptable as is, do *not* record its "Origno" on the Item Bias Review Sheet. Only items that should be revised or discarded should be recorded on the Item Bias Review Sheet.**

After you have completed your review, return *all* materials in the enclosed self-addressed envelope.

A Note about Test Security
It is important to note that these achievement test items are the property of the North Carolina Department of Public Instruction. If the items are not securely held, they will be useless to us. Do not copy the items; do not show them to anyone else; do not discuss their content with other people; and do keep them in a secure place when you are not reviewing them. Your help with security is essential to producing a test that is fair to all students.

Item Bias Review Sheet:Rejected Items

Name of Reviewer _____

Directions When reviewing the test items, keep in mind the following five questions:

1. Does the item contain any offensive gender, ethnic, regional, and/or ethnic content?
2. Does the item contain gender, ethnic, or cultural stereotyping?
3. Does the item contain activities that will be more familiar to one group than another?
4. Do the words in the item have a different meaning in one group than in another?
5. Could there be group differences in performance that are unrelated to proficiency in the content area?

If the answer is *Yes* to any of the questions, record the five-digit "Origno" below and check the appropriate column(s). You do not need to record items that are acceptable.

	Standard Failed						
Origno (Item #)	1	2	3	4	5	Revise	Reject

Published 1996

BEST COPY AVAILABLE